Supplementary Material

File S1 Glossary

The following glossary gathers the operational definitions adopted throughout the article. Each term is accompanied by a concise definition and the field bibliographic reference when the concept is inherited from prior literature. The terms are ordered by thematic affinity: general framework, rules R1–R6, Observatory profiles, dimensions D1–D4, and operational categories. Table S1 lists them.

Table S1. Glossary of the operational terms adopted throughout the article, with the field reference when the concept is inherited from published work.

Term	Definition
DPI	Digital Polarization Index. Operational framework for account classification and operation typification proposed in this article.
API	Application Programming Interface. Layer enabling programmatic access to platform data.
RT	Retweet (re-broadcast). Mechanism by which an account replicates another account message without modifying it.
Stance (DPI)	Proprietary DPI variable: directional position of a text toward the target case actor. Takes values Attacks, Defends, or No stance.
Platform Sentiment	Variable exported by some commercial monitoring tools: emotional polarity assigned by linguistic analysis to text in abstract, without reference to the target actor. Not used as DPI input; contrasted with Stance in subsection 4.2.5.2.
Side	Conceptual grouping of the seven operational categories into four: Trolls, Activists, Public opinion, and Neutral position.
Trolls (positive / negative)	Accounts satisfying at least one of the six rules R1–R6. The sign indicates dominant stance.
Activists (positive / negative)	Accounts with clear stance toward the actor, verifiable presence, low lexical aggressiveness, and at least some own content ($n \geq 3$ mentions).
Opinions (positive / negative)	Accounts with clear stance toward the actor without meeting verifiable presence or Troll criteria.
Neutral position	Accounts without dominant stance toward the actor.
Alphanumeric name pattern	Structural indicator of mass account generation: username ending in four or more consecutive digits (Cresci et al., 2017).
Verifiable presence	Binary indicator of authentic institutional presence: platform verification status. Followers are excluded as a manipulable signal in polarization operations.
No presence	Complement of verifiable presence: unverified account.
Textual replication	Proportion of a user mentions sharing identical text with other corpus accounts.
Pure amplifier	Account whose activity consists almost exclusively of retweeting without producing own content and without verifiable presence. Operationalizes rule R6 of the framework (Ng and Carley, 2023).
Reach operator	Functional type of operational account within the Troll class. Structural profile combining absence of institutional verification with high follower count and pure retweet behavior of coordinated text. Operationally: $pct_rt \geq 0.95 \wedge pct_coord \geq 0.95 \wedge followers \geq 5{,}000 \wedge \neg verified \wedge$ clear stance. Original DPI concept.
Coordinated text	Single text whose replication satisfies Eq. (1): at least 100 distinct accounts replicating it, together with either a velocity of at least 1 distinct account per minute or a burst of at least 30 replications within a fifteen-minute window.
Volume	Number of distinct accounts that have replicated a same text in the corpus.

Velocity	Distinct accounts per minute replicating a text over its activity window. Computed as vol divided by the duration in minutes of the window, with a floor of one minute.
Burst	Abnormal concentration of text replications in a fifteen-minute window (Ferrara et al., 2016).
Non-textual temporal coordination	Statistically abnormal synchrony of attacks without textual replication. Operationally: fifteen-minute window with at least thirty distinct accounts, the majority stance direction reaching 0.80 of the mentions of the window, and at least half of those mentions authored by accounts without verifiable presence. The three thresholds are empirical calibration of this framework.
Scale qualifier	Descriptor accompanying the typological label according to the total number of corpus mentions: low (< 500), standard (500–5,000), high (5,000–20,000), massive (≥ 20,000).
Intensity qualifier	Descriptor accompanying the typological label according to the D1 value: low (< 33%), medium (33–66%), high (≥ 66%).
Sapiens Observatory	Transversal DPI device that accumulates account classifications across the analyzed events, allowing identification of recurrent accounts and persistent operators.
PERSISTENT	Account appearing in at least three country events of the Sapiens Observatory without meeting the COORDINATED_OPERATOR criterion. Records sustained presence across events, not operational classification.
COORDINATED_OPERATOR	Account with Troll classification in at least two distinct country events, within the Sapiens Observatory.
RECURRENT	Account appearing in exactly two country events of the Sapiens Observatory without Troll classification in two or more of them. Records repeated presence across events without operational classification; an account appearing in three or more such events is PERSISTENT instead, and one classified as Troll in two or more is COORDINATED_OPERATOR.
OCCASIONAL	Account appearing in a single country event of the Sapiens Observatory. It is the residual profile of the cascade and is excluded from the operational core reported in Figure 8.
Trans-event persistent core	Set of accounts appearing with a Troll profile in a sustained majority of the analyzed events of a same country. Operationalizes the notion of reused digital infrastructure (Bradshaw, Bailey, and Howard, 2021; DiResta et al., 2021).
Operational sub-cores	Disjoint sets of coordinated operators operating on distinct frames within the same country, without account overlap.
Frame	Dominant interpretive framework around which a public conversation is structured. Operationally, the set of narratives sharing a common premise vis-à-vis the analyzed actor.
Mass replication	DPI typological modality: operation whose signature combines coordinated text detected by Eq. (1) with high D2 and D1 values.
Narrative saturation	DPI typological modality: operation whose signature combines high narrative concentration with non-textual temporal coordination or passive amplification.
Targeted amplification	DPI typological modality: operation whose signature combines the presence of recurrent operators of the Sapiens Observatory with a D1 value above the operational threshold.
No detectable operation	Fourth DPI typological label: corpus in which none of the three operational modalities is activated.
Botometer-Lite	Local reimplementation of the public Botometer logic on available structural signals (Varol et al., 2017; Yang and Menczer, 2024). Used as external reference in subsection 4.2.5.1.

POA	Public Opinion Analysis. Complementary module to the DPI, proposed to analyze organic content not framed in the operational frames of the lexicon through topic modeling.
Situational reading	DPI interpretive principle: the categories assigned by the model to an account describe its behavior within the corpus of a given event and do not constitute permanent attributes.

File S2. Operational thresholds

The following table gathers the detailed operational thresholds of the DPI framework, verified against the pipeline code. Values are organized into four blocks: chromatic thresholds of the four dimensions D1–D4, formal conditions of rules R1–R6, criteria of Equation (1) for coordinated text detection, and ranges of the two qualifiers accompanying the typological label. Table S2 reports them.

Table S2. Operational thresholds of the DPI framework, verified against the pipeline code and organized in four blocks: chromatic thresholds of the four dimensions, parameters of Equation (1), parameters of rules R1 to R6, and cut points of the scale and intensity qualifiers.

Block 1 — Chromatic thresholds by dimension	
D1 Troll intensity	Green 0–33% · Amber 33–66% · Red ≥ 66%
D2 Narrative concentration	Green 0–50% · Amber 50–80% · Red ≥ 80%
D3 Passive amplification	Green 0–25% · Amber 25–50% · Red ≥ 50%
D4 Temporal coordination	Green 0–10% · Amber 10–30% · Red ≥ 30%
Block 2 — Operational Troll rules	
R1 — Extreme alphanumeric pattern	pat_name ≥ 3 (name ends in ≥ 6 consecutive digits)
R2 — Moderate alphanumeric pattern without presence	pat_name == 2 (4–5 digits) ∧ ¬verified ∧ clear stance (otherwise the account is assigned to Neutral position)
R3 — Lexical aggressiveness with clear stance	pct_aggressive ≥ 0.30 ∧ clear stance ∧ ¬verified ∧ n ≥ 2
R4 — Repeated amplifier	n ≥ 3 ∧ pct_rt ≥ 0.95
R5 — High textual replication	rep_textual ≥ 0.70 ∧ n ≥ 2
R6 — Pure amplifier	n ≥ 1 ∧ pct_rt ≥ 0.95 ∧ ¬verified ∧ clear stance
Block 3 — Equation (1) for coordinated text	
Textual coordination	vol ≥ 100 ∧ (vel ≥ 1.0 ∨ burst ≥ 30)
Non-textual temporal coordination (15-minute window)	n_accounts ≥ 30 ∧ concentrated stance > 0.80 ∧ pct accounts without presence > 0.50
Block 4 — Qualifiers	
Scale qualifier (by n mentions)	Low: n < 500 · Standard: 500 ≤ n < 5,000 · High: 5,000 ≤ n < 20,000 · Massive: n ≥ 20,000
Intensity qualifier (derived from D1)	Low: D1 < 33% · Medium: 33% ≤ D1 < 66% · High: D1 ≥ 66%

The following tables document the aggressive lexicon on which rule R3 operates: the construction cascade from the source resource to the lexicon actually loaded by the pipeline, its categorical composition, the empirical distribution of the aggressive share over the study and control corpora, and the sensitivity of the aggregate Troll share to the two parameters of the rule.

Three categories are excluded in full: animals, words related to homosexuality, and words with potential negative connotations. The first and the third were discarded as noisy in political conversation, where animalized and ambiguous vocabulary is pervasive in non-aggressive registers; the second was discarded because its Spanish entries mix clinical and anatomical terminology with slurs, and the restrictive filter does not separate them reliably. The removal list carries 59 entries and reaches 18 forms that survive the cascade, among them "mentira", "falso", "pobre", "triste", "bandido" and "manta"; the remaining 41 are defensive and match nothing. The addition list carries 57 entries, 56 distinct after normalization and 53 not already present, and covers three families: strong Latin American insults absent from the Spanish distribution or present only as bigrams, despective political compounds, and regional forms.

The variable pct_aggressive is computed per account. The string submitted to matching is the normalized text produced at the stance stage: the mention lowercased and stripped of its leading retweet prefix, with diacritics preserved. That string is tokenized with the regular expression [a-z]+, which returns maximal runs of unaccented Latin letters and discards digits, punctuation, mention markers and hashtag markers. A mention is counted as aggressive when the intersection between its token set and the lexicon is non-empty. Matching is therefore whole-token and case-insensitive, with no substring matching, stemming or morphological expansion. The unit of count is the mention and not the token occurrence: a mention carrying five lexicon terms and a mention carrying one contribute equally. The variable is the ratio between the aggressive mentions of an account and its total mentions in the case corpus, and takes values in the closed interval from zero to one.

One property of this tokenization must be stated explicitly, because it bounds what the variable measures. The lexicon entries are diacritic-free by construction, since the build cascade folds them to their unaccented forms, whereas the mention text retains its diacritics. For the tokenizer an accented character is therefore not a letter but a token boundary, and an accented Spanish form is split rather than folded. The consequence runs in both directions. Aggressive forms carrying an accent are not recognized: across the eleven corpora this affects 799 mentions of "imbéciles", 217 of "vergüenza" and 195 of "patán", among others. Conversely, a fragment left by a split can coincide with a lexicon entry: 1,415 mentions are counted as aggressive because "artículo", "vehículo", "espectáculo" or "ridículo" yield the token "culo". The net effect on the results reported here was measured rather than assumed. Recomputing pct_aggressive under accent folding for all eleven corpora and reclassifying every account moves D1 by at most 0.2 percentage points in any single case, moves the aggregate Troll share of the study set from 76.52% to 76.55% and that of the control corpus from 34.15% to 34.11%, and changes no typological label and no intensity qualifier. The effect is bounded because the two directions of error partly offset each other and because R3 is the assigning rule for fewer than one in ten Troll accounts, as reported below.

The requirement of at least two mentions excludes accounts contributing a single mention to the corpus, for which pct_aggressive can only take the values zero or one and therefore carries no evidence of a sustained pattern. Its effect is measurable: 4,207 accounts of the study set satisfy the aggressiveness, stance and absence-of-presence conditions of R3 with a single mention, and 175 of them are classified as Troll by no other rule, so relaxing the requirement to one mention would raise the aggregate troll share from 76.52% to 77.14%. Under first-match attribution in the R1 to R6 order, R3 is the assigning rule for 2,056 of the 21,431 Troll accounts of the study set, that is 9.6%, and for 30 of the 992 of the control corpus, that is 3.0%; 183 of those 2,056 accounts satisfy no other troll rule, so the non-redundant contribution of R3 to the aggregate troll share is 0.65 percentage points. The rule is neither dominant nor ornamental: it isolates a behavioral signature, sustained aggressive vocabulary with unidirectional stance and without verifiable presence, that the structural rules do not capture.

The derived lexicon is calibrated but not exhaustive, and the direction of its error is known. The conservative level, the explicit stereotype flag and the four-character minimum trade recall for precision by construction, and the manual layer reflects the vocabulary of the corpora available when the lexicon was built rather than a systematic dialectal survey: the 53 added forms were triggered by two Ecuadorian corpora, fifty by EC/EV001 and seven by EC/EV003. Regional slang and emergent lexical formations that carry aggressive load in Latin American political conversation are consequently under-represented. Because unrecognized aggressive vocabulary lowers pct_aggressive, the bias is conservative: it produces false negatives in R3 and under-estimates rather than over-estimates the troll share. Lexical matching also models no context, so negation, quotation and irony are not distinguished, a limitation that supervised approaches to aggression and severity detection address at the cost of labeled training data [41].

Table S3. Construction of the aggressive lexicon from HurtLex ES.

Step	Criterion	Entries	Distinct forms
Source resource	HurtLex ES as distributed	5,006	3,354
Category filter	14 of the 17 categories retained	3,765	2,604
Level filter	level = conservative	1,698	1,087
Stereotype filter	stereotype = yes	618	481
Normalization	lowercase, NFD, combining marks stripped	-	458
Form filter	unigrams only, length >= 4 characters	-	435
Manual removal	stop list of 59 entries, 18 of them effective	-	417
Manual addition	regionalism list of 57 entries, 53 net new forms	-	470

Table S4. HurtLex categories and their contribution to the derived lexicon.

Code	Category	Included	Distinct forms contributed
ddp	Cognitive disabilities and diversity	Yes	213
dmc	Moral and behavioral defects	Yes	192
ps	Negative stereotypes and ethnic slurs	Yes	63
is	Social and economic disadvantage	Yes	25
rci	Locations and demonyms	Yes	10
pa	Professions and occupations	Yes	9
ddf	Physical disabilities and diversity	Yes	5
asf	Female genitalia	Yes	0
asm	Male genitalia	Yes	0
cds	Derogatory words	Yes	0
or	Plants	Yes	0
pr	Words related to prostitution	Yes	0
re	Felonies and words related to crime and immoral behavior	Yes	0
svp	Words related to the seven deadly sins	Yes	0
an	Animals	No	-
om	Words related to homosexuality	No	-
qas	Words with potential negative connotations	No	-

Table S5. Distribution of pct_aggressive and attribution of rule R3, by corpus. Accounts are the unique authors of the case corpus. The percentile column reports the position of the value 0.30 within the empirical distribution of pct_aggressive of that corpus, so a high percentile indicates a corpus in which few accounts reach the cut-off.

Country	Corpus	Accounts	pct_aggressive = 0	Accounts >= 0.30	Share >= 0.30	Percentile of 0.30	Trolls	Assigned by R3	R3 share of Trolls
EC	Noboa–Uribe	6,422	44.6%	3,460	53.9%	46.1	5,563	1,002	18.0%
EC	J. J. Neira	184	100.0%	0	0.0%	100.0	65	0	0.0%
EC	I. Vélez	2,288	97.2%	36	1.6%	98.4	2,072	15	0.7%
EC	C. A. Vera	821	39.7%	489	59.6%	40.4	603	78	12.9%
EC	Banco Pichincha	11,904	73.0%	2,068	17.4%	82.6	7,541	566	7.5%
DO	Yayo Sanz	31	100.0%	0	0.0%	100.0	2	0	0.0%
DO	Gloria Reyes	398	91.5%	23	5.8%	94.2	117	0	0.0%
MX	A. Armenta	3,899	55.1%	1,675	43.0%	57.0	3,696	390	10.6%
MX	F. Noroña	113	93.8%	7	6.2%	93.8	27	2	7.4%
MX	Clara Brugada	1,947	99.5%	8	0.4%	99.6	1,745	3	0.2%

-	Study set, pooled	28,007	67.4%	7,766	27.7%	72.3	21,431	2,056	9.6%
EC	Control corpus	2,905	79.8%	389	13.4%	86.6	992	30	3.0%

Table S6. Sensitivity of the aggregate troll share to the parameters of rule R3.

Parameter	Variant	Troll share, study set	Troll share, control corpus
Aggressiveness cut-off	>= 0.20	76.74%	34.25%
Aggressiveness cut-off	>= 0.25	76.66%	34.18%
Aggressiveness cut-off	>= 0.30 (adopted)	76.52%	34.15%
Aggressiveness cut-off	>= 0.40	76.28%	34.04%
Aggressiveness cut-off	>= 0.50	76.22%	34.01%
Aggressiveness cut-off	>= 0.60	76.03%	34.01%
Minimum mentions	n >= 1	77.14%	34.22%
Minimum mentions	n >= 2 (adopted)	76.52%	34.15%
Minimum mentions	n >= 3	76.30%	34.15%

File S3. Botometer-Lite specification

S.3.1. Scope and provenance

Botometer-Lite is a local reimplementation, written for this study, of the publicly documented structural markers exploited by the Botometer family of detectors [10]. It is not the Botometer model: it does not use its trained weights and does not query its service. The original classifies accounts from approximately one thousand one hundred fifty attributes obtained through the platform API, most of which—account age, timeline metadata, follower and friend graphs, per-tweet engagement—are unavailable under the API restriction described in subsection 2.1. The reimplementation retains the five markers that survive in a commercial monitoring export: username morphology, absence of verifiable presence, retweet ratio over own content, mention volume, and textual replication with other corpus accounts. The design follows the direction the field took after that restriction, toward detection and benchmarks built on the structural and behavioral signals that remain publicly accessible [8,9].

S.3.2. Features, weights and cut points

The score is the weighted sum of five feature scores, each a step function of one account-level variable. Table S7 states them as implemented.

Table S7. Botometer-Lite features, cut points and weights, transcribed from the pipeline implementation.

Feature	Variable	Cut points and feature score	Weight
F1. Username pattern	account handle	trailing run of 6 or more digits, 1.0; trailing run of 4 or 5 digits, 0.7; trailing run of 2 or 3 digits, 0.4; handle not ending in a digit but containing 5 or more digits and 8 or more characters, 0.6; otherwise 0.0, a handle ending in exactly one digit included	0.35
F2. Absence of verifiable presence	verification flag, follower count	1.0 if the account is unverified and reports fewer than 5,000 followers; otherwise 0.0	0.20

F3. Retweet ratio	share of retweets over the account's own mentions	0.95 or above, 1.0; 0.80 or above, 0.7; 0.60 or above, 0.4; otherwise 0.0	0.25
F4. Volume anomaly	mentions in the corpus	50 or more, 0.9; exactly 1, 0.5; otherwise 0.0	0.10
F5. Textual replication	replication index against other corpus accounts	0.7 or above, 1.0; 0.5 or above, 0.6; otherwise 0.0	0.10

The five weights add to 1.000, so the normalization the code applies is a constant divisor and the score is bounded in the zero-to-one interval by construction; the result is rounded to three decimals. An account is labeled bot when the score is greater than or equal to 0.500, the criterion stated in subsection 4.2.5.1. Two discrepancies between the module documentation and the implemented code are recorded here for exactness: the documentation states the decision criterion as a strict inequality while the implemented comparison is the non-strict one, a difference that subsection S.3.4 shows to be consequential; and the documentation enumerates four features in use while five are scored, textual replication being absent from the enumeration and present in the code. The five signals are far from equally active over the ten study corpora: F2 activates in 94.4% of the 28,007 account records, F3 reaches its first cut in 86.8%, F1 takes a non-zero value in 35.9%, the single-mention branch of F4 applies to 65.3% while its hyperactivity branch applies to eleven accounts, and F5 reaches its first cut in thirty-seven accounts, 0.1% of the set. Three features carry the score in practice.

S.3.3. Absence of training and statistical calibration

The weights and the cut points were not estimated from data. No labeled account sample was used, no supervised fitting was performed, no threshold was optimized against a validation set, and no performance metric—precision, recall, area under the curve—was computed for this reference, since computing one would require ground truth the study does not have. The weights encode a bibliographic judgment on the relative diagnostic value of the five markers, and the cut points are round values placed where the descriptive literature locates suspicion; the term "calibrated" applied to Botometer-Lite in subsection 4.2.5.1 and in the module documentation designates this bibliographic anchoring and not statistical calibration. Botometer-Lite is therefore a literature-informed heuristic and not a replication: it does not reproduce the performance characteristics of the original Botometer model, and no claim about its accuracy, its false-positive rate or its comparability with published Botometer benchmarks is made or implied in this article. The caution the field has documented applies in full: bot attribution from public structural signals is fragile, contested in its construct validity, and systematically prone to false positives in political corpora [17,29,30].

S.3.4. Mass on the decision boundary

Over the pooled set of 28,007 account records of the ten study corpora, 9,584 accounts, 34.2% of the set, receive a score of exactly 0.500 and are labeled bot by the non-strict comparison. All of them share a single profile: unverified, fewer than 5,000 followers, retweet ratio at or above 0.95, one mention in the corpus, no scoring digit pattern in the handle, and textual replication below the first cut of F5. Their score is the sum of 0.200, 0.250 and 0.050, and nothing else. They constitute

51.7% of the 18,538 accounts the reference labels bot. Applying instead the strict inequality stated in the module documentation leaves 8,954 bots, lowers raw concordance with the DPI Troll classification from 73.8% to 48.3%, and lowers the aggregate Cohen kappa from 0.368 to 0.132. The agreement reported in subsection 4.2.5.1 therefore depends substantially on the direction in which one arbitrary tie is broken. The tied population is moreover the single-mention pure amplifier profile that the framework reaches under an independent criterion: 8,371 of these 9,584 accounts, 87.3%, also hold a clear stance toward the target actor and are classified as Trolls, almost all of them through rule R6, while the remaining 1,213 hold no stance and are classified as Neutral position.

S.3.5. Divergence in the presence criterion

Feature F2 declares absence of verifiable presence when the account is unverified and reports fewer than 5,000 followers. The DPI defines verifiable presence by platform verification status alone and excludes follower count explicitly, on the ground that follower count is a purchasable and therefore manipulable signal; the rules that invoke presence read it from the verification flag only, and follower count, although carried in the account table, enters no rule. The reference and the framework compared against it consequently do not share the definition of the variable that carries 0.20 of the reference score. Over the pooled set, 1,328 accounts, 4.7%, are unverified yet report 5,000 followers or more: the reference credits them with presence, the framework does not. Aligning F2 with the DPI definition by removing the follower clause changes the bot label of 626 accounts, 2.2% of the set, raises raw concordance from 73.8% to 75.5%, and raises the aggregate kappa from 0.368 to 0.392. The magnitude is small, the direction is not neutral: a fraction of the measured disagreement originates in a definitional divergence internal to the comparison rather than in a difference of detection behavior.

S.3.6. Consequences for the reported agreement

Three properties of the reference bound what the kappa of 0.368 can support: it was never validated against labeled accounts, a third of the population sits exactly on its decision boundary, and one of its five features is defined against the framework's own criterion. The unit of aggregation adds a fourth qualification: the 28,007 records are account-corpus pairs corresponding to 25,274 distinct handles, since an account recurring across events enters once per corpus. Under these conditions a fair agreement is compatible with the complementarity reading offered in subsection 4.2.5.1 and equally compatible with the reading that attributes part of the discordance to the arbitrariness of the reference. Both are reported as readings. The contrast establishes that the DPI output is not a relabeling of a structural bot detector, which is the descriptive claim the evidence supports; it does not establish the accuracy of either classification, and the hypothesis contrast of subsection 4.2.4 does not depend on it.

S.3.7. Reproducibility

The scoring function is deterministic and depends only on account-level fields present in the per-case classification files, so it can be applied to any corpus processed by the pipeline. The archived script requires path adaptation before execution: its root directory is an absolute path inherited from the environment in which the comparison was first run, it enumerates three of the ten study cases under folder names that predate the current naming convention, it reads the account file from a directory one level above its current location, and it writes its summary outside the repository tree; scoring function and case loop share module scope, so importing the module executes the loop. Executed unmodified it stops at the first read with a file-not-found error. The figures reported in subsection 4.2.5.1 reproduce exactly when the unmodified scoring function is applied to the ten current per-case account files.

File S4. Sensitivity analysis of the operational thresholds

S.4.1. Reimplementation and its validation

The sensitivity analysis was conducted on an independent reimplementation of the DPI rather than on the production pipeline, so that the reported panel remains the one produced by the audited pipeline at nominal thresholds. The reimplementation reconstructs the two threshold-bearing stages of the framework: the detection of coordinated texts of Equation (1) and the four-dimensional panel with its typological labeling, including the recomputation of the Sapiens Observatory recurrence profiles under each parameter setting.

Before any threshold was moved, the reimplementation was required to reproduce the official panel exactly at nominal values. Agreement was verified at five independent levels of granularity over the eleven corpora of the study, comprising 64,582 mentions, 30,912 unique accounts and 11,557 unique texts. Table S8 reports the outcome. Agreement is exact at every level, including the per-account Troll assignment for all 30,912 accounts and the four inputs of Equation (1) for all 11,557 unique texts. Exact agreement establishes the reimplementation as a faithful reconstruction at the operating point; away from it, the results below are properties of the thresholds under the parameterization documented in section S.4.2.

Three implementation details are recorded for completeness. First, the field reported in the panel output as the number of coordinated texts is a Boolean indicator of the existence of at least one such text, not a count; the actual counts of coordinated texts are 3, 1, 16, 2 and 2 for EC/EV001, EC/EV003, EC/EV005, MX/EV001 and the control corpus respectively, and zero elsewhere. Second, the recurrence profiles of the Sapiens Observatory depend on the troll classification and were therefore recomputed under every parameter setting rather than held fixed, excluding the control corpus from the accumulation. At nominal thresholds the recomputed profiles and the stored ones differ by at most nineteen accounts in a single corpus and yield identical typological labels for all eleven corpora, since the labeling rule reads only whether the intersection is non-empty. Third, the digit-pattern score of R1 and R2 carries a secondary branch for handles with five or

more scattered digits, which is held fixed when the digit bounds are varied; the variation of those two thresholds therefore acts on the trailing-digit criterion alone.

S.4.2. Design of the sweep

Twenty-four thresholds were varied: the three parameters of Equation (1) plus its burst window; the four parameters of the D4 detector; the ten parameters of rules R1 to R6; the shared clear-stance cut-off; the retweet ratio governing the passive amplifier set of D3; and the four pairs of chromatic cut-offs of D1 to D4, each pair scaled jointly. Every threshold took four alternative values at -50%, -25%, +25% and +50% of nominal, one threshold at a time with the remaining twenty-three at nominal, giving ninety-six configurations. Ratios were capped at 1.00 where the nominal value scaled above unity, which affects the retweet ratios of R4, R6 and D3 at +25% and +50%, the D4 stance concentration at +50%, and the textual replication of R5 at +50%; those cells are reported at the cap and the cap is noted. Discrete mention counts were compared against the unrounded scaled value, so that, for example, a nominal n at least 2 varied to +25% becomes an effective requirement of n at least 3. The eleven corpora comprise the ten study cases and the organic control corpus, which is not part of the validation set and does not enter the Observatory.

Beyond the four-point envelope, each threshold was additionally scanned over a fine grid to locate the exact interval within which every typological label of the eleven corpora is identical to the nominal outcome. Those intervals are the stability margins reported in Table S9.

S.4.3. Threshold-by-threshold stability

Table S9 reports, for each threshold, the number of corpora whose typological label changes at each of the four variations, the interval of values over which all eleven labels are unchanged, and the largest displacement induced in D1 and D4. Two aggregate results frame the table. First, of the 880 corpus-configuration cells produced by the eighty non-chromatic configurations, 849 (96.5%) carry the same set of typological labels as the nominal run; the 31 changes (3.5%) concentrate in six thresholds. Second, D2 is invariant under every threshold varied, because narrative concentration is computed from the case lexicon and does not read any of the parameters in Table S9; and D3 is likewise invariant, for a structural reason developed in section S.4.4.

The dimensions move asymmetrically. D1 is displaced by up to 68.2 percentage points, D4 by up to 100.0 percentage points, and D2 and D3 not at all. Table S13 reports the per-corpus envelope: over the eighty configurations, D1 for EC/EV001 ranges from 38.7% to 90.8% against a nominal 86.6%, and D4 for DO/EV001 ranges from 0.0% to 100.0% against a nominal 100.0%. The dimensional envelopes are therefore wide while the labels are narrow, which is the central asymmetry of the framework: the typological assignment rests on the existence of qualifying structures, not on the precise magnitude of the dimensions.

S.4.4. Critical and inert thresholds

The distinction that answers the stability question directly is between thresholds whose variation reassigns typological labels and thresholds whose variation does not.

Six thresholds are critical. The minimum burst of Equation (1) reassigns two corpora at -50% and two at +50%, and holds all labels only over the interval 25 to 31 replications; it is the tightest margin in the framework. The burst window of Equation (1) is stable over 13.5 to 18.5 minutes. The minimum number of unique accounts of the D4 detector is stable over 24 to 31 accounts and reassigns four corpora at +50%. The D4 window is stable over 11.5 to 21.0 minutes. The D4 stance concentration is stable over 0.52 to 0.98. The lower digit bound of R2 reassigns one corpus at -50%, when it admits handles ending in two digits, and is otherwise stable up to at least eight digits.

Three of these six govern D4, and their effect runs through the existence of at least one qualifying window rather than through the magnitude of D4. Raising the minimum account count from 30 to 32 removes the single window of DO/EV001, and with it both the Narrative saturation label and the whole of its D4 value; the corpus falls to the No detectable operation label, the only configurations in the entire sweep where that label is ever activated. This is a property of the smallest corpus of the set, of 31 mentions, and not of the detector in general.

Fourteen of the twenty non-chromatic thresholds are label-inert over the full envelope, but inertness on labels does not imply inertness on dimensions, and the distinction is worth stating separately. Three cases arise.

The first is structural inertness. The share of accounts without verifiable presence required by the D4 detector, nominally 0.50, changes nothing anywhere between 0.20 and 0.80, because in each of the seventy coordinated regions reported by the detector that share lies between 0.946 and 1.000; the criterion is satisfied with a margin of roughly a factor of two and does not discriminate. The retweet ratio of D3 is inert for a stronger reason: D3 is computed over accounts with exactly one mention, for which the retweet ratio can only take the values 0.0 or 1.0 and the stance proportions can only take the values 0.0 or 1.0. Any cut-off strictly inside the unit interval selects the identical set of accounts, so D3 is invariant not by calibration but by construction. The same argument makes D3 invariant to the clear-stance cut-off.

The second is inertness with genuine dimensional movement. The minimum mention count of R6, nominally 1, is the most consequential threshold of the framework for the account-level classification: raising it to an effective 2 reclassifies 36.84% of all 30,912 accounts, moves D1 by up to 68.2 percentage points, changes the intensity qualifier of seven corpora and the D1 color band of seven corpora, and lowers the Spearman correlation of the D1 ordering with the nominal ordering to 0.806, its minimum over the whole sweep. It nevertheless reassigns no typological label in any corpus. The threshold is also asymmetric, since 1 is the minimum admissible value and no downward variation is possible; the entire effect is one-sided. R1, the mention count of R4, and the clear-stance cut-off displace D1 by 9.3, 5.5 and 6.9 percentage points respectively without reassigning labels.

The third is inertness on both counts. The minimum volume and minimum velocity of Equation (1), the aggressiveness threshold and mention count of R3, the replication threshold and mention count of R5, and the retweet ratios of R4 and R6 leave every label unchanged and displace D1 by at most 5.5 percentage points. The minimum volume in particular is label-inert from at most 20 to at least 250 accounts, a range that contains the 50 to 200 interval recorded in the methodological documentation. The minimum velocity is inert over the envelope and well beyond it upward, but its stability is bounded below at 0.35 replications per minute: at 0.30 and below the burst criterion ceases to be the binding one and two corpora acquire the Mass replication label.

The chromatic cut-offs form a separate class. By construction they cannot reassign a typological label, since no labeling rule reads them; their variation only moves corpora between color bands. They are nonetheless the least robust element of the panel: scaling the D1 pair by +50% moves seven of eleven corpora across a band boundary, and the D3 pair at -50% moves five. This is a presentational fragility, not a classificatory one, and it argues for reading the bands as a communication aid rather than as a finding.

S.4.5. Rationale of each threshold

Table S10 states the basis of every threshold. Three provenances are distinguished and are not presented as equivalent. Seven thresholds derive from published operational criteria and are cited accordingly: the minimum velocity of Equation (1) [1,11], the minimum burst [12], the digit patterns of R1 and R2 [38], the retweet ratios of R4 and R6 [9], and the textual replication threshold of R5 [13]. Two are field conventions with no single attributable source: the simple-majority clear-stance cut-off, and the fifteen-minute window shared by the burst of Equation (1) and the D4 detector, whose length is nonetheless consistent with the interval over which false cascades exhibit their peaks [12].

The remainder rest on empirical calibration over the study corpora and on no bibliographic authority, and are stated as such without qualification. The minimum volume of 100 accounts in Equation (1) was calibrated on the Noboa–Uribe corpus. The minimum of 30 unique accounts of the D4 detector was calibrated on the Noboa–Uribe and Irene Vélez corpora. The minimum mention count of R6 was calibrated on the Irene Vélez corpus, where single-use amplifier accounts constitute the operational signature. The aggressiveness threshold of R3 combines calibration with the toxicity literature built on the HurtLex lexicon [35]. The stance concentration and no-presence share of the D4 detector are operational heuristics chosen to require unidirectionality and a majority of accounts without institutional presence, with no external calibration; the sensitivity results in section S.4.4 show that the second of the two is not doing discriminative work at any value in a wide interval around its nominal setting. The minimum mention counts of R3, R4 and R5 are design decisions intended to suppress single-mention false positives, a known failure mode of behavioral detectors [17]. The chromatic cut-offs rest on internal calibration over

the eleven corpora, with the sole exception of the 66% red boundary of D1, whose justification is given in subsection 4.1.

S.4.6. Should the thresholds depend on context?

The framework operates almost entirely on proportions, which are dimensionless and therefore transportable across corpora of different size. One threshold is not: the minimum volume of Equation (1) requires at least 100 distinct accounts replicating the same text, an absolute count, and it is consequently sensitive to corpus size in a way no other parameter is.

Table S11 quantifies the effect. For each corpus it reports the share of the corpus that a single text would have to capture in order to reach 100 replications. The spread is three orders of magnitude. In the Banco Pichincha corpus of 32,381 mentions a text needs 0.3% of the corpus; in the Noboa–Uribe corpus, 1.0%; in the control corpus, 1.3%. At the other end, the JJ Neira corpus of 197 mentions requires 50.8%, the F. Noroña corpus of 130 mentions requires 76.9%, and the Yayo Sanz corpus of 31 mentions would require 322.6%, which is to say that Equation (1) cannot fire in that corpus under any circumstance whatsoever. Three of the eleven corpora are effectively outside the reach of the Mass replication label, and their largest observed replication reaches 97, 11 and 1 accounts respectively. The consequence is that the absence of the Mass replication label in a small corpus is not evidence of the absence of mass replication; it is partly an artefact of the absolute threshold.

The empirical impact of this asymmetry on the reported panel is nil, since the sweep shows the minimum volume to be label-inert from at most 20 to at least 250 accounts. The limitation is structural rather than manifest: it bites only on corpora that are already too small for the criterion, and it would bite harder on any future corpus in the same size range.

A corpus-relative formulation was tested directly, replacing the fixed minimum by a share of the corpus. Table S12 reports the outcome. A purely relative rule is unstable in both directions: at 0.5% of the corpus it converts three corpora to Mass replication, including the Yayo Sanz corpus whose largest text was replicated by a single account, and at 5% it removes the label from the Banco Pichincha and control corpora. A hybrid rule with an absolute floor, requiring the maximum of 30 accounts and 1% of the corpus, reproduces every typological label of the eleven corpora exactly while removing the size dependence for corpora above 3,000 mentions. The same holds for a floor of 30 and 2%.

The answer to whether the thresholds should depend on context is therefore differentiated. The proportional thresholds should not, and the evidence in Table S9 supports leaving them fixed. The minimum volume of Equation (1) should, and a hybrid floor-plus-share formulation is label-equivalent on the present corpus set while eliminating a real dependence on corpus size. The framework retains the nominal absolute value in this study so that the reported panel remains the one produced by the audited pipeline, and the hybrid formulation is recorded as the specific change that the evidence supports for subsequent versions of the index.

A second, weaker form of context dependence deserves recording. The Targeted amplification label requires a non-empty intersection with the accounts profiled as PERSISTENT or COORDINATED_OPERATOR in the Sapiens Observatory, and the size of that reference population differs by country: 1,249, 721 and 918 matching accounts in the Banco Pichincha, Noboa–Uribe and Irene Vélez corpora against 27, 17 and 11 in the Mexican ones and a single account in each Dominican corpus. The label is thus easier to activate where the Observatory has accumulated more history. This is a property of the accumulation and not of a threshold, and it is not addressed by any variation in Table S9.

S.4.7. Effect on the conclusions of the manuscript

No conclusion reported in section 4 is overturned by any variation within the envelope tested, although individual corpora do change label under six of the twenty-four thresholds, as Table S9 records.

The empirical differentiation of the three active modalities, which is the success criterion of hypothesis H1, survives all eighty non-chromatic configurations: Mass replication, Narrative saturation and Targeted amplification each remain activated in at least one of the ten study cases in 80 of 80 configurations. The counts move within bounded ranges, from 4 cases at nominal to between 2 and 6 for Mass replication, from 7 to between 3 and 8 for Narrative saturation, and from 9 to between 9 and 10 for Targeted amplification. The number of distinct label combinations across the ten cases ranges from 3 to 5 against 4 at nominal, so the combinatorial variety on which the differentiation argument depends is never collapsed.

The No detectable operation label, reported in subsection 4.2.4 as absent from the ten study cases, is activated in only 3 of the 80 configurations and always in the same corpus, the 31-mention Yayo Sanz case, and only when the D4 account minimum is raised above 31 or its window is halved. The design caveat recorded in subsection 4.2.4 is therefore unaffected.

The relative reading of the panel is preserved. The Spearman correlation between the nominal ordering of the ten cases by D1 and the ordering under each variation has a median of 1.000 and a minimum of 0.806, the minimum occurring under the R6 mention-count variation discussed in section S.4.4.

The organic control corpus behaves as the study corpora do under perturbation: it retains its nominal labels in 79 of the 80 configurations, its D1 ranges from 19.0% to 40.4% against a nominal 34.1%, and the number of study cases falling below it in D1 ranges from 2 to 4 against 3 at nominal. Its nominal panel is reported in the main text and is not a result of this file.

The single element of the framework that the analysis identifies as genuinely fragile is presentational. The chromatic bands, which carry no classificatory weight, are the only component that moves substantially under proportional variation of its own cut-offs, and they should be read accordingly.

S.4.8. Limitations of the sensitivity analysis

Five limitations bound the reach of the exercise. First, the design is one-at-a-time: every configuration moves a single threshold and holds the other twenty-three at nominal, so interactions are not explored, and the joint variation of two critical thresholds may displace labels that neither displaces alone. Second, the envelope is bounded at 50% in each direction, and three thresholds are already shown to break outside it, so the stability margins of Table S9 are statements about a neighbourhood and not about the whole parameter space. Third, the reimplementation was validated against the pipeline at nominal values only; its fidelity away from that point rests on the correspondence of the parameterization with the operational code, not on an independent check. Fourth, the invariance of D2 is a property of the sweep design rather than a robustness result: narrative concentration depends on the narrative lexicon of each case, which is a calibrated artefact and was not varied here, so D2 is untested rather than shown to be stable. Fifth, the eleven corpora belong to three countries and one platform, and every margin reported here is a property of that corpus set; the smallest corpora in particular drive most of the label reassignments observed, so the margins should be expected to narrow as corpora of intermediate size are added.

Table S8. Validation of the reimplementation against the pipeline output at nominal thresholds.

Validation level	Unit compared	Units	Exact matches	Agreement
Panel dimensions D1–D4	Corpus	11	11	100.0%
Typological labels	Corpus	11	11	100.0%
Coordinated temporal regions detected	Corpus	11	11	100.0%
Troll / non-Troll assignment	Account	30,912	30,912	100.0%
Equation (1) inputs and flag (volume, duration, velocity, burst, coordination flag)	Unique text	11,557	11,557	100.0%

Table S9. Threshold-by-threshold stability of the typological labels over the eleven corpora.

Threshold	Nominal	-50%	-25%	+25%	+50%	Stable interval	Max |ΔD1|	Max |ΔD4|	Class
Eq. (1) minimum volume	100 accounts	0	0	0	0	≤20 to ≥250	0.0	0.0	Inert
Eq. (1) minimum velocity	1.0 per min	0	0	0	0	0.35 to ≥5.0	0.0	0.0	Inert
Eq. (1) minimum burst	30 replications	2	1	2	2	25 to 31	0.0	0.0	Critical

Eq. (1) burst window	15 min	2	2	0	1	13.5 to 18.5	0.0	0.0	Critical
D4 minimum unique accounts	30 accounts	1	1	3	4	24 to 31	0.0	100.0	Critical
D4 stance concentration	0.80	1	0	1	1†	0.52 to 0.98	0.0	67.0	Critical
D4 share without presence	0.50	0	0	0	0	≤0.10 to ≥1.00	0.0	0.0	Inert
D4 window	15 min	4	1	0	1	11.5 to 21.0	0.0	100.0	Critical
R1 trailing digits	6	0	0	0	0	≤3 to ≥10	9.3	0.0	Sensitive
R2 trailing digits, lower bound	4	1	0	0	0	3 to ≥8	22.5	0.0	Critical
R3 lexical aggressiveness	0.30	0	0	0	0	≤0.05 to 1.00	0.7	0.0	Inert
R3 minimum mentions	2	0	0	0	0	≤1 to ≥5	4.4	0.0	Inert
R4 retweet ratio	0.95	0	0	0†	0†	≤0.30 to 1.00	4.7	0.0	Sensitive
R4 minimum mentions	3	0	0	0	0	≤1 to ≥7	5.5	0.0	Sensitive
R5 textual replication	0.70	0	0	0	0†	≤0.20 to 1.00	0.3	0.0	Inert
R5 minimum mentions	2	0	0	0	0	≤1 to ≥5	0.5	0.0	Inert
R6 retweet ratio	0.95	0	0	0†	0†	≤0.30 to 1.00	2.4	0.0	Inert
R6 minimum mentions	1	0	0	0	0	1 to 2	68.2	0.0	Sensitive
Clear-stance cut-off (R2, R3, R6, D3)	0.50	0	0	0	0	≤0.10 to ≥0.95	6.9	0.0	Sensitive
D3 retweet ratio	0.95	0	0	0†	0†	≤0.30 to 1.00	0.0	0.0	Inert (structural)
D1 color cut-offs	33 / 66	0	0	0	0	not label-bearing	0.0	0.0	Presentational
D2 color cut-offs	50 / 80	0	0	0	0	not label-bearing	0.0	0.0	Presentational

D3 color cut-offs	25 / 50	0	0	0	0	not label-bearing	0.0	0.0	Presentational
D4 color cut-offs	10 / 30	0	0	0	0	not label-bearing	0.0	0.0	Presentational

Table S10. Rationale of each operational threshold.

Threshold	Nominal	Basis
Eq. (1) minimum volume	100 accounts	Empirical calibration on the Noboa–Uribe corpus. No bibliographic source.
Eq. (1) minimum velocity	1.0 replication per minute	Published operational criterion [1,11]
Eq. (1) minimum burst	30 replications	Published operational criterion [12]
Eq. (1) burst window	15 min	Field convention; window in which false cascades exhibit their peaks [12]
D4 window	15 min	Coherence with the burst window [12]
D4 minimum unique accounts	30 accounts	Empirical calibration on the Noboa–Uribe and Irene Vélez corpora. No bibliographic source.
D4 stance concentration	0.80	Operational heuristic requiring clear unidirectionality. No calibration and no source.
D4 share without presence	0.50	Operational heuristic requiring a majority without presence. No calibration and no source.
R1 trailing digits	≥ 6	Mass account generation pattern [38]
R2 trailing digits	4 to 5	Same pattern under a relaxed bound [38]
R3 lexical aggressiveness	≥ 0.30	Empirical calibration combined with the toxicity literature built on HurtLex [35]
R3 minimum mentions	≥ 2	Design decision against single-mention false positives [17]
R4 retweet ratio	≥ 0.95	Published operational criterion [9]
R4 minimum mentions	≥ 3	Design decision: recurrence requirement
R5 textual replication	≥ 0.70	Published operational criterion [13]
R5 minimum mentions	≥ 2	Design decision against single-mention false positives
R6 retweet ratio	≥ 0.95	Published operational criterion, differentiated from R4 [9]
R6 minimum mentions	≥ 1	Empirical calibration on the Irene Vélez corpus to capture single-use amplifier accounts. No bibliographic source.
Clear-stance cut-off	> 0.50	Field convention: simple majority over mentions with clear stance
D3 retweet ratio	≥ 0.95	Coherence with R6. Structurally non-discriminative, see section S.4.4.
D1 color cut-offs	33 / 66	Internal calibration; the 66% boundary is justified in subsection 4.1
D2, D3, D4 color cut-offs	50 / 80, 25 / 50, 10 / 30	Internal calibration over the eleven corpora. No bibliographic source.

Table S11. Corpus share required to reach the absolute minimum volume of Equation (1).

Corpus	Mentions	Unique texts	Share of corpus needed for 100 replications	Reachable	Largest observed replication	Texts above 100
EC · Banco Pichincha	32,381	5,477	0.3%	Yes	877	62
EC · Noboa–Uribe	9,708	176	1.0%	Yes	2,249	15
EC · control corpus	7,959	3,180	1.3%	Yes	259	5
MX · Armenta	5,765	943	1.7%	Yes	1,590	5
EC · Irene Vélez	4,062	394	2.5%	Yes	755	10
MX · Clara Brugada	2,399	226	4.2%	Yes	563	6
EC · C. A. Vera	1,031	226	9.7%	Yes	457	3
DO · Gloria Reyes	919	764	10.9%	Yes	43	0
EC · JJ Neira	197	25	50.8%	Yes	97	0
MX · F. Noroña	130	115	76.9%	Yes	11	0
DO · Yayo Sanz	31	31	322.6%	No	1	0

Table S12. Typological outcome under corpus-relative formulations of the minimum volume.

Rule for the minimum volume	Corpora changing label	Changes
vol ≥ 0.5% of corpus	3	EC · JJ Neira, MX · F. Noroña and DO · Yayo Sanz gain Mass replication
vol ≥ 1% of corpus	1	DO · Yayo Sanz gains Mass replication
vol ≥ 2% of corpus	1	DO · Yayo Sanz gains Mass replication
vol ≥ 5% of corpus	2	EC · Banco Pichincha and the control corpus lose Mass replication
vol ≥ max(30, 1% of corpus)	0	none
vol ≥ max(30, 2% of corpus)	0	none
vol ≥ 100 (nominal)	—	reference

Table S13. Per-corpus envelope of the dimensions over the eighty non-chromatic configurations.

Corpus	D1 nominal	D1 range	D3 nominal	D3 range	D4 nominal	D4 range	Distinct label sets	Configurations with nominal labels
EC · Noboa–Uribe	86.6	38.7–90.8	39.4	39.4–39.4	42.0	14.4–95.5	1	100.0%
EC · JJ Neira	35.3	11.4–44.6	28.4	28.4–28.4	0.0	0.0–0.0	1	100.0%
EC · Irene Vélez	90.6	40.5–93.0	33.9	33.9–33.9	33.6	1.0–77.2	2	95.0%
EC · C. A. Vera	73.4	31.9–76.0	44.6	44.6–44.6	0.0	0.0–21.9	3	95.0%

EC · Banco Pichincha	63.3	38.1–70.2	11.4	11.4–11.4	27.9	0.0–94.9	2	97.5%
DO · Yayo Sanz	6.5	3.2–29.0	0.0	0.0–0.0	100.0	0.0–100.0	3	95.0%
DO · Gloria Reyes	29.4	10.3–36.7	9.6	9.6–9.6	10.0	0.0–13.4	2	96.2%
MX · Armenta	94.8	45.5–95.9	48.3	48.3–48.3	10.3	0.0–42.0	3	90.0%
MX · F. Noroña	23.9	13.3–41.6	0.0	0.0–0.0	0.0	0.0–0.0	1	100.0%
MX · Clara Brugada	89.6	21.4–90.4	61.4	61.4–61.4	22.1	0.0–80.9	3	93.8%
EC · control corpus	34.1	19.0–40.4	6.3	6.3–6.3	0.0	0.0–33.6	2	98.8%

File S5. Lexical pattern load per study case

The stance of a mention is decided by the case lexicon specified in subsection 3.1 and File S11. Table S14 reports, for each study case, the single lexical pattern that decides the largest number of mentions, the share of the corpus that pattern decides on its own, and the share of the corpus matching two or more substantive patterns. In six of the ten cases the most load-bearing pattern decides fewer than one tenth of the mentions on its own; the exception that matters is the largest corpus of the set, where a neutral entity name decides almost half of them.

Table S14. Lexical pattern load per study case. Coverage is the share of mentions receiving a directional stance. Patterns are reproduced as they are stored in the lexicon, in lower case and without diacritics, which is the matching normalization described in subsection 3.1.

Case	Mentions	Lexical coverage	Most load-bearing pattern	Mentions decided by that pattern alone (%)	Mentions matching two or more patterns (%)
MX/EV002	130	1.0	doralacensuradora	87.7	12.3
EC/EV005	32,381	0.7012	el banco pichincha	44.7	24.2
MX/EV001	5,765	1.0	chingatumadrearmenta	23.1	76.9
MX/EV003	2,399	0.9921	ajolotizacion	19.2	80.1
DO/EV002	919	0.3787	quesecreehomero	8.3	19.2
EC/EV003	4,062	0.9993	irene velez	6.8	91.9
DO/EV001	31	1.0	campanapoliticadesdeaduana	0.0	100.0
EC/EV004	1,031	0.8555	hijobobo	0.0	100.0
EC/EV001	9,708	0.8519	niega	0.0	72.3
EC/EV002	197	0.3299	que sabe neira	0.0	33.0

File S6. Organic control corpora

These four corpora fall outside the entry condition stated in subsection 3.2 and are discussed in subsection 5.3. Table S15 reports the panel each control returns against the range observed over the ten study cases, Table S16 the outcome label by label, and Table S17 the diffusion morphology of the fourteen corpora.

Table S15. Control corpora and the panel returned by the DPI, against the range observed over the ten study cases. Labels: MR Mass replication, NS Narrative saturation, TA Targeted amplification.

Corpus	Event	Date	Mentions	Accounts	D1	D2	D3	D4	Registered operators	Labels
EC/CTRL001	Association football derby	12 Jul 2026	7,959	2,905	34.1	10.0	6.3	0.0	105	MR, TA
EC/CTRL002	Association football, cup tie	20 Aug 2026	4,234	1,253	30.6	38.7	4.1	0.0	23	TA
EC/CTRL003	Cycling stage victory	25 Jul 2026	9,949	3,038	56.6	29.9	8.5	5.6	129	MR, NS, TA
DO/CTRL001	Music concert	21 Feb 2026	906	654	18.0	62.8	7.5	0.0	0	none
Study cases (10), range	political, trending	—	31 – 32,381	31 – 11,904	6.5 – 94.8	10.5 – 100.0	0.0 – 61.4	0.0 – 100.0	1 – 1,249	all ten labeled

Table S16. Label-level outcome over the four control corpora.

Label	Activates on	False positives	Condition that drives the activation
Mass replication	2 of 4	2	Equation (1) imposes no condition on the origin of the replicated text
Narrative saturation	1 of 4	1	stance concentration above threshold within a fifteen minute window
Targeted amplification	3 of 4	3	size of the accumulated national record, not the corpus analyzed

Table S17. Diffusion morphology across the fourteen corpora. Origin is the account whose post is being replicated for retweets and the author otherwise. Gini is computed over the distribution of mentions across origins.

Corpus	Retweets %	Distinct origins	Top ten origins %	Gini of origins	Single mention accounts %	Retweet only accounts %
EC/CTRL001	61.9	1,375	29.4	0.734	54.6	54.3
EC/CTRL002	61.4	715	42.1	0.734	51.0	50.0
EC/CTRL003	83.8	1,000	46.9	0.831	52.5	72.9
DO/CTRL001	39.5	416	30.6	0.484	81.5	38.1
Study cases, median	85.9	160.5	71.7	0.843	79.7	86.1
Study cases, range	0.0 – 98.5	21 – 2,826	16.9 – 93.4	0.000 – 0.900	55.0 – 100.0	0.0 – 98.4

File S7. Robustness with respect to the calibration corpus

The dimensional thresholds documented in subsection 4.1 and File S2 were verified by sensitivity analysis on the Ecuadorian corpus EC/EV001 Noboa–Uribe, which also belongs to the set of ten validation cases. This file isolates the consequences of that overlap by recomputing the aggregate results of section 4 over the nine corpora that did not participate in threshold verification. The check is a leave-one-out exercise on the evidence set, not a modification of the framework: the six rules R1–R6, the four dimensions, Equation (1), the coordination-window criteria, the chromatic cut-offs, and the label assignment rules of Table 4 are held at the values reported in File S2,

and the values published in Table 5 remain the nominal-threshold output of the pipeline. The thresholds are not re-derived from the nine corpora, so the check bounds the influence of the calibration corpus on the reported results and does not establish threshold optimality.

Two caveats delimit the reach of the exercise. The nine corpora are disjoint from threshold verification but not statistically independent of the withdrawn one, since the four remaining Ecuadorian events share account populations with it: of the 1,971 accounts of the Ecuadorian operational core, 949 appear in EC/EV001 and 557 lose core status when it is withdrawn, leaving 1,414 accounts, 71.7% of the original core, whose status does not depend on the calibration corpus. In addition, the Sapiens Observatory that supplies the recurrent-operator count of the Targeted amplification rule is built inclusively of the case being labeled, so that count is not independent of the case in either evidence set. Table S18 compares the two evidence sets, and Table S19 reports which condition of the coordination window governs each corpus.

The reduced set comprises 46,915 mentions and 21,585 unique accounts, 82.9% and 77.1% of the original volume. Each of the three operational modalities is still activated: Mass replication in three cases, Narrative saturation in six, Targeted amplification in eight. Ecuador continues to activate the three modalities, Mexico the three, and the Dominican Republic two, and the cross-country coverage of each modality is unchanged: Narrative saturation and Targeted amplification are each present in at least one case of the three countries, while Mass replication remains confined to the Ecuadorian and Mexican corpora. The amplitude of the four dimensions is unchanged, because all eight extremes belong to cases outside the calibration corpus: D1 between 6.5% (DO/EV001 Yayo Sanz) and 94.8% (MX/EV001 A. Armenta), D2 between 10.5% (EC/EV005 Banco Pichincha) and 100.0%, a value reached by four cases, D3 between 0.0% and 61.4% (MX/EV003 Clara Brugada), and D4 between 0.0% and 100.0% (DO/EV001). The medians descend by between 2.8 and 10.4 percentage points, which registers the withdrawal of a high-signature case and not a contraction of range. The differential activation that the success criterion of H1 requires therefore holds on evidence disjoint from calibration, and the separability criterion of H2 is not met on the reduced set either. Hypothesis H1 retains its status and its caveat without modification: the No detectable operation label is activated in none of the nine remaining cases.

The two external contrasts of subsection 4.2.5 are unaffected in direction and shift only marginally in magnitude. Against Botometer-Lite, the aggregate agreement over n = 21,585 unique accounts moves to a Cohen kappa of 0.398, which remains in the fair range [40], with raw concordance of 73.8%, 75.4% of DPI Trolls also flagged by the automatic detector, and 87.2% of the detector bots simultaneously DPI Trolls. Against Platform Sentiment, the aggregate agreement over n = 46,871 mentions moves to a Cohen kappa of -0.027, farther from equivalence than the ten-case value, with diagonal concordance of 45.7%. Both are point estimates without interval estimation, so the displacement is reported as a direction and not as a tested difference.

Neither the complementarity reading nor the null-agreement reading depends on the calibration corpus.

Hypothesis H3 is supported on the reduced set, with the Ecuadorian magnitudes reduced in proportion to the corpus withdrawn. The Sapiens Observatory operational core falls from 2,052 to 1,495 accounts; the Mexican core of twenty-seven COORDINATED_OPERATOR accounts and the Dominican core of one are unchanged, since the profiles are country-scoped. In Ecuador the core falls from 1,971 to 1,414 accounts, the COORDINATED_OPERATOR set from 1,345 to 968, and the PERSISTENT set from 48 to 24. The 968 surviving operators are a strict subset of the original 1,345, and the reduced Ecuadorian core is likewise a strict subset of the complete one: the withdrawal removes evidence without producing any operator that the complete set did not already identify. The Ecuadorian intra-country nucleus survives the withdrawal. The fourteen accounts present in the five Ecuadorian corpora are contained by construction in the intersection of the four remaining ones; the informative result is that this intersection holds twenty-four accounts and that twenty-three of them receive the COORDINATED_OPERATOR profile from an Observatory rebuilt without the calibration corpus, so the nucleus is recoverable from corpora that calibrated no threshold. The negative trans-country observation is preserved but loses evidential weight: the single account that appeared in events of two distinct countries belongs to the withdrawn corpus, so on the reduced set the intersection between the account populations of the three countries is empty in the three pairs by construction rather than by independent confirmation.

Two findings of the check qualify the framework rather than confirm it, and both are developed below. First, the condition of stance concentration inside the coordination window, which the design treats as the behavioral discriminator of D4, governs the value of the dimension in only one of the nine remaining corpora; in six of them the governing condition is the count of unique accounts, in the remaining two no fifteen-minute interval reaches thirty mentions at all, and the requirement that most accounts of a window lack verifiable presence is inert over the whole study set. Second, the invariance of the typological label under perturbation of the D4 thresholds holds for every case with a corpus above one thousand mentions and is not guaranteed below that magnitude: of the four corpora under that size, the two Dominican ones lose a label and the other two are invariant only because D4 is null in every configuration tested.

The distribution of typological modalities preserves its shape: four cases activated three simultaneous labels and three still do, two cases activated two labels and two still do, four cases activated a single label and four still do, and no case receives the No detectable operation label. The ranking of modalities by frequency is unchanged, Targeted amplification above Narrative saturation above Mass replication, and the empirical association between Mass replication and corpus magnitude survives: the three cases that activate it hold 4,062, 5,765, and 32,381 mentions, all above the four-thousand-mention level reported in subsection 4.2.3. The polarity pattern also survives: eight of

the nine cases exhibit unilateral hostile polarity and DO/EV002 Gloria Reyes remains the only structural exception.

The Ecuadorian recurrent-operator counts that feed the Targeted amplification rule decline when the Observatory is rebuilt without the withdrawn corpus, from 119 to 102 in EC/EV002, from 918 to 788 in EC/EV003, from 459 to 423 in EC/EV004, and from 1,249 to 959 in EC/EV005, and remain above zero in every case. No typological label of any of the nine cases changes as a result of the recomputation.

The contraction of the Ecuadorian PERSISTENT set from forty-eight to twenty-four accounts is in part an artifact of holding the criterion at its absolute value. The profile requires presence across at least three country events, so on four accumulated events it demands three of four rather than three of five, a proportionally stricter bar. Eighteen of the original forty-eight accounts retain the profile, thirty descend to RECURRENT because the withdrawn corpus supplied one of their three appearances, and six accounts previously classified COORDINATED_OPERATOR reach PERSISTENT because that corpus supplied one of their two troll classifications.

A window of fifteen minutes is retained by the D4 detector when it contains at least thirty unique accounts, when the larger of its "Attacks" and "Defends" counts reaches 0.80 of its mentions, and when at least half of its mentions come from accounts without verifiable presence. Over the ten corpora, 41,809 windows reach the minimum of thirty mentions and none of them fails the presence condition; over the reduced set the figure is 32,613 windows and, again, zero failures. The condition contributes no selectivity on any corpus of this study, because the account populations of the analyzed events are overwhelmingly non-verified and the constraint is satisfied trivially.

Of the 32,613 candidate windows of the reduced set, 874, or 2.7%, fail the count of unique accounts, and 24,675, or 75.7%, fail the stance concentration. The aggregate proportion is misleading and the per-corpus reading is the informative one, since the stance concentration removes windows in only two of the nine corpora, EC/EV005 Banco Pichincha with 24,450 windows and DO/EV002 Gloria Reyes with 225. Only in the first does it govern the value of the dimension: displacing the concentration threshold from 0.70 to 0.90 moves D4 in EC/EV005 from 42.5% to 8.8% and leaves D4 numerically invariant in the other eight corpora, DO/EV002 included, whose four retained windows survive the whole interval. In six of the nine corpora the governing condition is the count of unique accounts, and in EC/EV002 J. J. Neira and MX/EV002 F. Noroña no fifteen-minute window reaches thirty mentions at any threshold tested, so D4 is settled by conversational density before any behavioral criterion applies.

The complementary reading identifies the specific cost of having verified thresholds on a single corpus. In EC/EV001 the unique-account condition is close to inert: displacing its threshold from twenty-six to forty accounts moves D4 from 43.0% to 41.3%, a range of 1.7 percentage points, while displacing the concentration threshold from 0.70 to 0.90 moves it from 70.9% to 29.8%. In the corpora that did not calibrate, the ordering is reversed, and the same displacement of the unique-account threshold moves D4 from 46.0% to 11.3% in EC/EV003 I. Vélez, from

41.5% to 1.9% in MX/EV003 Clara Brugada, from 16.3% to 0.0% in MX/EV001 A. Armenta, from 13.4% to 0.0% in DO/EV002, and from 100.0% to 0.0% in DO/EV001. The calibration corpus is therefore the one least able to probe the condition that governs D4 in six of the nine remaining cases, and the sensitivity analysis reported in subsection 4.1 measured the elasticity of the dimension with respect to the condition operative there rather than with respect to the condition operative elsewhere.

The stability of the typological label was measured over two grids centered on the nominal values: unique accounts in {28, 30, 32} crossed with stance concentration in {0.78, 0.80, 0.82}, and unique accounts in {25, 30, 35} crossed with stance concentration in {0.75, 0.80, 0.85}. The account threshold is displaced by roughly seven percent in the first grid and roughly seventeen percent in the second, and the concentration threshold by roughly three and six percent. Both grids yield the same result. The label is invariant in seven of the nine cases across all nine configurations. The two exceptions are the two Dominican corpora. DO/EV001 Yayo Sanz, with thirty-one mentions and thirty-one accounts, loses its only label at a unique-account threshold of thirty-two, since the requirement then exceeds the size of the account universe, and the case moves to No detectable operation. DO/EV002 Gloria Reyes, with 919 mentions and 398 accounts, loses Narrative saturation at the same threshold, because none of its 277 candidate windows retains thirty-two distinct accounts, and keeps Targeted amplification. Every case with a corpus above one thousand mentions is label-invariant across both grids.

The boundary is a property of scale rather than of the rules, and the invariance below it is weaker than the count suggests. The D4 detector requires an absolute count of thirty coordinating accounts inside a fifteen-minute window, so a corpus whose account population, or whose per-window occupancy, is of that order cannot separate detection from non-detection under a perturbation of the threshold. Of the four corpora below one thousand mentions, two change label and the other two, EC/EV002 J. J. Neira with 197 mentions and MX/EV002 F. Noroña with 130, are invariant only in the degenerate sense that D4 is null in every configuration tested and their label does not depend on the D4 thresholds at all. The robustness against scale documented in subsection 4.1 therefore holds for the numerical value of the dimensions, which are proportions and remain interpretable at any corpus size, and does not extend to the stability of the typological label in corpora of a few hundred accounts. On the calibration corpus alone the panel and the label are invariant in every configuration of both grids, which is the reason a verification restricted to that corpus could not surface this boundary.

The following results are sustained on evidence disjoint from threshold calibration: the differential activation of the three operational modalities across the reduced set; the amplitude of the four dimensions, whose eight extremes lie outside the calibration corpus; the cross-country coverage of the three modalities; the association between Mass replication and corpus magnitude; the unilateral hostile polarity pattern and its single structural exception; the fair

complementarity with Botometer-Lite, with the aggregate Cohen kappa moving from 0.368 to 0.398; the null agreement between DPI Stance and Platform Sentiment, with the aggregate Cohen kappa moving from -0.015 to -0.027; and the trans-event persistence of operators in the three countries, including the Ecuadorian nucleus of fourteen accounts, all of which are recoverable from the four non-calibrating corpora.

The following results are attenuated or qualified. The Ecuadorian magnitudes reported above fall by between 28.0% and 50.0% without any operator being created that the full set did not already contain, and the contraction of the PERSISTENT set is partly the mechanical effect of holding an absolute criterion over fewer accumulated events. The negative trans-country observation becomes an empty intersection in the three country pairs, but by construction, since the only account that connected two countries belonged to the withdrawn corpus. The reading of stance concentration as the behavioral discriminator of D4 holds in only one of the nine non-calibrating corpora and is therefore a property of high-volume events rather than a general property of the dimension, while the requirement of non-verifiable presence is inert across the whole study set. The invariance of the typological label under threshold perturbation holds for every corpus above one thousand mentions and is not guaranteed below that magnitude, which sets an explicit lower bound on the corpus size at which the label, as distinct from the panel, should be read as a stable output of the framework.

Table S18. Aggregate results of the study with and without the calibration corpus EC/EV001 Noboa–Uribe. The nine-case column recomputes every quantity over the corpora that did not participate in threshold verification, holding the rules, the dimensions and the thresholds at their nominal values. Observatory profiles are country-scoped, so withdrawing an Ecuadorian corpus leaves the Mexican and Dominican records unchanged.

Quantity	Ten cases	Nine cases	Change
Mentions	56,623	46,915	-17.1%
Unique accounts, case by case	28,007	21,585	-22.9%
Unique identities after within-country deduplication	25,274	19,801	-21.7%
Cases activating Mass replication	4	3	-1
Cases activating Narrative saturation	7	6	-1
Cases activating Targeted amplification	9	8	-1
Cases activating No detectable operation	0	0	0
Cases with three / two / one simultaneous labels	4 / 2 / 4	3 / 2 / 4	-1 / 0 / 0
D1 range (%)	6.5 to 94.8	6.5 to 94.8	unchanged
D2 range (%)	10.5 to 100.0	10.5 to 100.0	unchanged
D3 range (%)	0.0 to 61.4	0.0 to 61.4	unchanged
D4 range (%)	0.0 to 100.0	0.0 to 100.0	unchanged
D1 / D2 / D3 / D4 median (%)	68.4 / 87.2 / 31.2 / 16.2	63.3 / 76.8 / 28.4 / 10.3	-5.1 / -10.4 / -2.8 / -5.9
DPI vs Botometer-Lite, unique accounts	28,007	21,585	-22.9%
DPI vs Botometer-Lite, raw concordance	73.8%	73.8%	0.0 pp

DPI vs Botometer-Lite, Cohen kappa	0.368	0.398	+0.030
DPI Trolls also flagged Bot	76.2%	75.4%	-0.8 pp
Botometer bots also DPI Trolls	88.0%	87.2%	-0.8 pp
Stance vs Platform Sentiment, mentions	56,579	46,871	-17.2%
Stance vs Platform Sentiment, diagonal concordance	45.2%	45.7%	+0.5 pp
Stance vs Platform Sentiment, Cohen kappa	-0.015	-0.027	-0.012
Observatory operational core, three countries	2,052	1,495	-27.1%
Ecuador operational core	1,971	1,414	-28.3%
Ecuador COORDINATED_OPERATOR	1,345	968	-28.0%
Ecuador PERSISTENT	48	24	-50.0%
Mexico COORDINATED_OPERATOR	27	27	unchanged
Dominican Republic COORDINATED_OPERATOR	1	1	unchanged
Accounts present in every Ecuadorian corpus	14 (of five)	24 (of four)	nucleus preserved
Accounts appearing in events of two countries	1	0	empty intersection

Table S19. Governing condition of the D4 coordination window over the nine non-calibrating corpora. A candidate window is a fifteen-minute interval containing at least thirty mentions. The three conditions are applied in the order shown, so a window that fails the account count is not tested against the later ones. The governing condition is the criterion whose displacement changes the value of the dimension. In DO/EV002 the stance concentration rejects most candidate windows without governing the value, which responds only to the unique-account threshold; in DO/EV001 no candidate window is rejected at nominal thresholds, and the dimension collapses when the unique-account threshold passes the size of the account universe. The calibration corpus is reported in the last row for comparison and is not included in the total.

Case	Candidate windows	Fail account count	Fail stance concentration	Fail presence	Windows retained	D4 (%)	Governing condition
EC/EV002 J. J. Neira	0	0	0	0	0	0.0	Window density
EC/EV003 I. Vélez	1,124	355	0	0	769	33.6	Unique-account count
EC/EV004 C. A. Vera	2	2	0	0	0	0.0	Unique-account count
EC/EV005 Banco Pichincha	30,679	387	24,450	0	5,842	27.9	Stance concentration
DO/EV001 Yayo Sanz	2	0	0	0	2	100.0	Unique-account count
DO/EV002 Gloria Reyes	277	48	225	0	4	10.0	Unique-account count
MX/EV001 A. Armenta	253	39	0	0	214	10.3	Unique-account count
MX/EV002 F. Noroña	0	0	0	0	0	0.0	Window density
MX/EV003 Clara Brugada	276	43	0	0	233	22.1	Unique-account count
Nine-case total	32,613	874 (2.7%)	24,675 (75.7%)	0 (0.0%)	7,064		

EC/EV001 Noboa–Uribe (calibration)	9,196	34	5,814	0	3,348	42.0	Stance concentration

File S8. Per-country dimensional panels

The three panels below compare the four DPI dimensions across the cases of a single country. They are reported here rather than in subsection 4.2.2 because they restate, in graphical form, values that Table 5 already carries, so that the running text keeps the comparison in tabular form and the graphical version remains available without interrupting it.

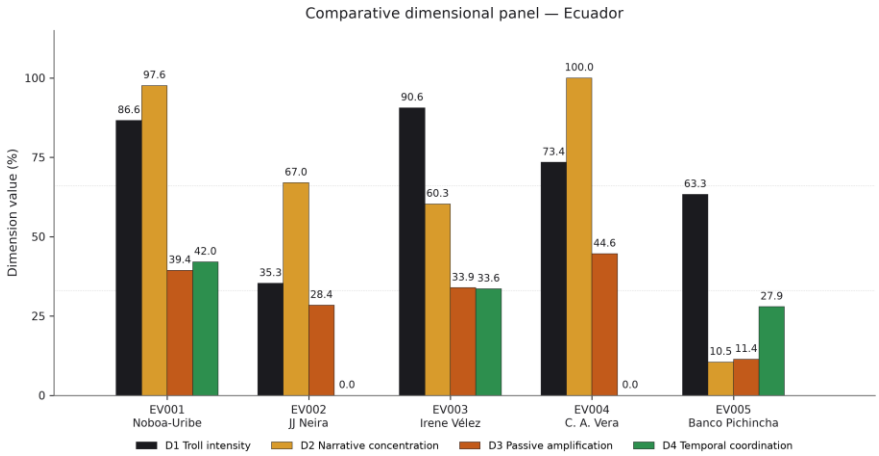


Figure S1. Compared dimensional panel of the five Ecuadorian cases. Each case is reported with its four DPI dimensions (D1 Troll intensity, D2 narrative concentration, D3 passive amplification, D4 temporal coordination). Horizontal dotted lines at 33% and 66% mark the chromatic thresholds of the framework. The panel allows comparison of dimensional profiles across cases from the same country.

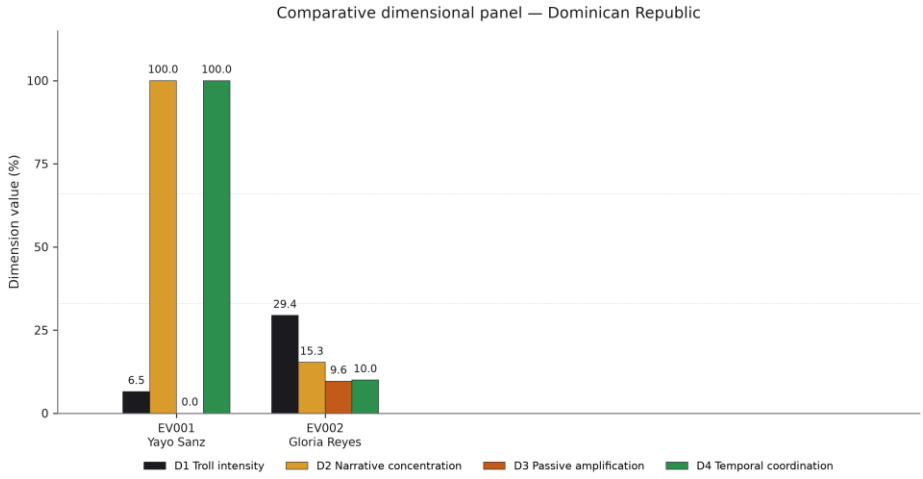


Figure S2. Compared dimensional panel of the two Dominican cases. Each case is reported with its four DPI dimensions (D1 Troll intensity, D2 narrative concentration, D3 passive amplification, D4 temporal coordination). Horizontal dotted lines at 33% and 66% mark the chromatic thresholds of the framework.

The panel allows comparison of dimensional profiles across cases from the same country.

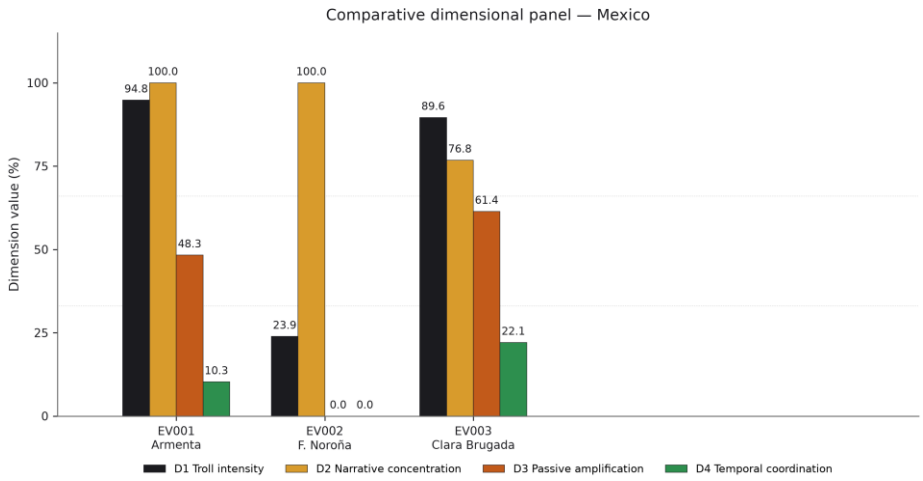


Figure S3. Compared dimensional panel of the three Mexican cases. Each case is reported with its four DPI dimensions (D1 Troll intensity, D2 narrative concentration, D3 passive amplification, D4 temporal coordination). Horizontal dotted lines at 33% and 66% mark the chromatic thresholds of the framework. The panel allows comparison of dimensional profiles across cases from the same country.

File S9. Divergence between the DPI Stance and the Platform Sentiment

This file develops the divergence reported in subsection 4.2.5.2: the composition of the two families of divergence types, worked examples of each, the structural properties that qualify the proportions, and the full typology in Table S20.

Two families organize seven of the eight types, and a residual type collects the remainder. The first family covers eight thousand six hundred eighty-one mentions, 28.0% of the discordance and 15.3% of the corpus, and gathers divergences of anchoring: the affect the platform scores is not anchored to the target actor by the DPI Stance. Unanchored affect is the larger component, with six thousand four hundred forty-four mentions. Its boundary is exact rather than interpretive: the text contains zero anti patterns and zero pro patterns of the case lexicon, so the DPI assigns "No stance" by construction, while the platform scores an emotional charge that the text does carry. Two situations fall inside this boundary and the measurements separate them. In 6.0% of these mentions the text does not name the target actor at all and the affect belongs to a lateral object —another institution, an adversary, a commercial entity that shares a name token with the actor— which is divergence of directionality in the strict sense. In the remaining 94.0% the text names the actor or its account handle and the emotional charge is plausibly directed at it, but the wording falls outside the case lexicon; these mentions record a coverage limit of the DPI Stance rather than a defect of the platform variable, and they are the empirical measure of the classifier error margin that subsection

5.2 declares. Adversary redirection, the second component, covers two thousand two hundred thirty-seven mentions in which the DPI assigns "Defends" and the platform assigns negative or neutral sentiment, because the defense of the actor is executed as an attack on the accuser. Its quantitative profile is unmistakable: 97.1% of these mentions name a third-party account, 84.8% carry a link, the median length is one hundred sixteen words —the longest of the eight types, two and a half times the corpus median of forty-five— and only 3.7% contain aggressive lexicon. The defense is long, documented, addressed to a third party, and lexically restrained.

The second family covers twenty-two thousand one hundred eighty-five mentions, 71.5% of the discordance and 39.2% of the corpus, and gathers divergences of attack vehicle: both variables point at the same object, but the aggression or the defense does not travel in affect-bearing prose, which is the only substrate the Platform Sentiment classifier scores. Affect-free derision is the single largest type of the entire discordance, with nine thousand two hundred thirty-five mentions, 29.8% of it. These texts contain no token of the aggressive lexicon, have a median length of twenty-five words, and 74.8% of them carry a link, so that the derisive payload sits in an image, a video, or a coined label —a nickname, a suffixed neologism formed on a local referent, a mock slogan— and not in scoreable words. Documentary framing covers six thousand three hundred fifty-three mentions in which the attack or the defense is stated in reportorial or institutional register, with 89.8% carrying hashtags and 95.4% being retweets of media or official accounts. Ironic courtesy framing covers two thousand six hundred eight mentions in which an attack is wrapped in a greeting, a congratulation, or an expression of gratitude; 64.3% contain aggressive lexicon and 68.3% contain emoji, and the platform reads the phatic surface. Hashtag-borne aggression covers one thousand six hundred fifty-two mentions in which the aggressive payload sits inside hashtag tokens that the classifier does not segment as lexical content; the running prose is affectively flat, the median length is twelve words, and 86.3% name a third-party account. Unmarked invective covers two thousand three hundred thirty-seven mentions that contain aggressive lexicon in running prose and still receive neutral sentiment; this is the only type in which the divergence is attributable to the platform classifier alone rather than to the vehicle of the attack, and it is also the most concentrated of the eight, with seventy-nine unique texts behind two thousand three hundred thirty-seven mentions. Outside both families, a residual type of one hundred sixty-six mentions, 0.5% of the discordance and twelve unique texts, gathers texts with an exact tie between anti and pro patterns.

Illustrative renderings of the types, translated from the original Spanish and with all account handles and third-party identities elided, are the following. Unanchored affect, lateral variant: "[A rival commentator collective] are a digital virus that infects the networks with nothing but hatred and discredit", classified "No stance" and scored negative, in a corpus whose target actor the text never names. Unanchored affect, actor-naming variant: "Nothing works, not [the bank], not the civil registry, not the public hospitals; is this really the

new country?", classified "No stance" because no case pattern matches, and scored negative. Adversary redirection: "THE KING OF FAKE NEWS: while [a neighboring country] faces four terrorist attacks in less than forty-eight hours, [a foreign head of state] resorted to a photograph three years old", classified "Defends" and scored negative. Documentary framing: "ATTENTION | continuous failures at [the institution]; while the regulator sets its eye on the savings and credit cooperatives and forces them to close, what crown does [the institution] wear that it is neither investigated nor reviewed nor sanctioned?", classified "Attacks" and scored neutral. Hashtag-borne aggression: "Who joins this petition? [obscene hashtag naming the actor] together with your gag-law hashtag; no to censorship", classified "Attacks" and scored neutral. Ironic courtesy framing: "Good morning, community; ready to snap our fingers at the despot [obscenity] [actor]? Let us go for it", classified "Attacks" and scored positive. Unmarked invective: "Look at the traitor to the homeland, [a foreign politician], who said he had not met [the actor], and the photographs of the meeting appeared, where they would have agreed to raise the tariffs", classified "Attacks" and scored neutral. Affect-free derision: "The new slogan of [the institution]", followed by an image, and "the [neologism]ization of the city came with an aquatic ecosystem included", both classified "Attacks" and scored neutral.

Three structural properties of the discordance qualify the reading of the proportions. The first is replication: 87.2% of the discordant mentions are retweets, and the thirty-one thousand thirty-two mentions reduce to four thousand six hundred sixty-three unique texts, of which twenty-three account for half of the total discordance and one hundred thirty-three for four fifths. Weighted at the level of the unique text rather than the mention, the type shares reorder: Unanchored affect rises from 20.8% to 37.5%, Affect-free derision holds at 24.3%, Hashtag-borne aggression rises from 5.3% to 20.2%, and the two types that depend on massively replicated single texts fall —Documentary framing from 20.5% to 8.1% and Unmarked invective from 7.5% to 1.7%. The mention-level proportions describe the discordance as the analyst encounters it in a corpus; the text-level proportions describe the repertoire of divergent forms. Both readings support the same conclusion and neither depends on a single case, but the effective number of independent observations behind the discordance is the number of texts and not the number of mentions. The second property is that the two families are not equally available in every corpus: anchoring divergence requires that the discordance contain mentions with "Defends" Stance or with "No stance" and no matched pattern. It is material only in the three corpora that contain the "Defends" level — EC/EV001, EC/EV005, and DO/EV002— where it reaches 12.7%, 51.2%, and 64.2% of the discordance respectively; it is residual in EC/EV002 and EC/EV003, with one and three mentions; and it is absent in the remaining five corpora, where the whole discordance falls in the second family or, in the case of EC/EV004, partly in the residual type, which contributes one hundred forty-eight of its one hundred sixty-six mentions. The third property is that the ten corpora are written in Spanish and scored by one family of commercial sentiment classifiers,

so the volumetric weight of each vehicle is a property of these corpora and not a general parameter; the aggressive lexicon that supports two of the eight types is likewise calibrated for Spanish.

Table S20. Typology of the divergence between the DPI Stance and the Platform Sentiment, over the 31,032 discordant mentions of the aggregated cross-country matrix (n = 56,579). The eight types are mutually exclusive and exhaustive: each mention receives exactly one type through a fixed precedence over the eleven mention-level measurements. "Family A" gathers divergences of anchoring (the affect scored by the platform is not anchored to the target actor by the DPI Stance); "Family B" gathers divergences of attack vehicle (the aggression or the defense does not travel in affect-bearing prose); "Family C" is the residual tie. The columns for aggressive lexicon, hashtags, link, and third-party handle report the percentage of mentions of the type in which the feature is present; the aggressive lexicon is the calibrated resource in Spanish used by rule R3 [35], and a third-party handle is any account handle other than the retweeted author and the target actor. Unique-text counts are computed within each type; they sum to 4,664 against the 4,663 distinct discordant texts of the corpus because one text is split between two types by the Platform Sentiment of its mentions.

Family	Divergence type	Contingency cells	Mentions	% of divergence	% of corpus	Unique texts	Median words	Aggressive lexicon	Hashtags	Link	Third-party handle
A	Unanchored affect	negative x "No stance"; positive x "No stance"	6,444	20.8	11.4	1,748	37	17.0%	18.0%	54.1%	32.2%
A	Adversary redirection	negative x "Defends"; neutral x "Defends"	2,237	7.2	4.0	200	116	3.7%	7.7%	84.8%	97.1%
B	Affect-free derision	neutral x "Attacks"; positive x "Attacks"	9,235	29.8	16.3	1,134	25	0.0%	34.4%	74.8%	11.8%
B	Documentary framing	neutral x "Attacks"; neutral and negative x "Defends"	6,353	20.5	11.2	376	51	37.0%	89.8%	51.6%	30.4%
B	Ironic courtesy framing	positive x "Attacks"; neutral x "Attacks"	2,608	8.4	4.6	172	32	64.3%	74.4%	92.9%	70.4%
B	Unmarked invective	neutral x "Attacks"	2,337	7.5	4.1	79	70	100.0%	3.3%	18.5%	3.3%
B	Hashtag-borne aggression	neutral x "Attacks"; positive x "Attacks"	1,652	5.3	2.9	943	12	4.1%	100.0%	50.7%	86.3%

C	Balanced-frame tie	negative x "No stance"	166	0.5	0.3	12	14	0.6%	89.2%	90.4%	11.4%
A	Anchoring subtotal		8,681	28.0	15.3	1,948					
B	Attack-vehicle subtotal		22,185	71.5	39.2	2,704					
C	Residual subtotal		166	0.5	0.3	12					

The distinctive signature of each type is the following. Unanchored affect: zero anti and zero pro patterns of the case lexicon, of which 6.0% are lateral, in that the actor is not named, and 94.0% name the actor but fall outside lexicon coverage. Adversary redirection: the longest texts of the eight types, with the defense executed as an attack on the accuser. Affect-free derision: no aggressive token, with the payload in an image, a video or a coined label. Documentary framing: reportorial or institutional register, 95.4% of it retweets. Ironic courtesy framing: a greeting, congratulation or gratitude formula wrapping an insult, 68.3% of it carrying emoji. Unmarked invective: aggressive lexicon in running prose scored neutral, at 29.6 mentions per unique text. Hashtag-borne aggression: payload confined to hashtag tokens, with affectively flat prose. Balanced-frame tie: an exact tie between anti and pro patterns.

File S10. Matched within-corpus control

This file reports the design and the full outcome of the matched contrast summarized in subsection 4.2.6: the sampling scheme, the circularity audit that separates rule inputs from variables no rule reads, the geometry of the separation, the behavior of the organic control corpus under the same design, and Table S21.

For each of the ten case corpora, and for the first of the organic control corpora, the association football derby of 12 July 2026 (7,959 mentions, 2,905 accounts), 500 matched draws were taken from each arm, each draw of size equal to the smaller group: the smaller arm enters whole and the larger one is subsampled without replacement, which in six of the nine retained case corpora means subsampling the Troll arm. Draws were compared with the two-sided Mann-Whitney U test [42] and summarized by Cliff's delta [43], a rank effect size bounded between -1 and 1, and by the overlapping coefficient OVL, the shared area of the two histograms on a common support. Reported values are medians over the 500 draws. One corpus (Dominican Republic EV001; 31 accounts, 2 Trolls) yields a two-account arm and is excluded, leaving nine case corpora and 6,270 matched pairs per draw against 992 pairs in the control corpus. Across the ten case corpora the classifier labels 21,431 of 28,007 accounts as Trolls; in the control corpus, 992 of 2,905.

The result must be read with its circularity declared in advance. Six of the nine variables compared are quantities the rules threshold: mentions contributed, retweet share, intra-account textual repetition, aggressive-lexicon share, handle digit-pattern score and platform verification. Dominant Stance is a seventh rule input, read by R2, R3

and R6, and is not among the nine. Any separation on these is tautological in part. Three variables are read by no surviving classification rule and carry whatever external validity the contrast has: the share of an account's mentions belonging to texts flagged as coordinated by Equation (1), follower count, which is excluded from the operational definition of presence, and hashtags per mention. The classifier binds the coordination and hashtag fields but no rule branch references them, since the two rules that once used them were withdrawn from the operative model. Coordination and follower count do enter the pipeline elsewhere, coordination in the case-level typology and both in an informative sub-classification column, but neither reaches the account-level label.

Among rule inputs, only the handle digit-pattern score separates the two groups in all nine case corpora, and its median effect is small (delta = 0.181). Retweet share separates in seven of nine, with the largest median effect of the panel (delta = 0.492). The remaining rule inputs barely move: intra-account textual repetition separates in none, platform verification in two, mentions contributed in four with a median effect of 0.005, and aggressive-lexicon share in five with a median effect of zero. This is a property of the label rather than a failure of the test. The label is a union of rules and each rule fires on a distinct minority: pure amplification supplies 56.2% of all Trolls, handle morphology 23.5%, repeat pure retweeting 10.7%, the aggressive lexicon 9.6%, and textual replication 0.05%. The dominant rule also imposes no effective threshold on volume, since its requirement of at least one mention is satisfied by every account in the corpus. Even the tautological half of the contrast is therefore partial: no single variable characterizes the category.

Among the variables no rule reads, follower count is lower for Trolls in all nine case corpora (median delta = -0.102, range -0.345 to -0.012), reaching significance in five, and coordinated-text share is higher in each of the four corpora where coordinated text exists at all, up to delta = 0.693. Hashtags per mention are unstable in sign (-0.749 to 0.785) and are not diagnostic. The less circular contrast restricts the Troll arm to the accounts whose label is decided by the handle rules R1 and R2, which read the digit pattern of the name string and, for R2, the absence of platform verification and a dominant Stance, but no volume, retweet, repetition or lexical behavior. Because the hierarchy evaluates the handle rules first, this arm is not behaviorally clean: 85.6% of its accounts also satisfy at least one of the behavioral Troll rules, so its higher retweet share (median delta = 0.241) cannot be read as out-of-rule evidence. The follower deficit can: these accounts have fewer followers than matched non-Trolls in nine of nine corpora (median delta = -0.324), and when the arm is further restricted to handle-flagged accounts that satisfy no behavioral rule the deficit persists in all eight corpora where that subset reaches ten accounts (median delta = -0.303, significant only in the largest), while their retweet share stops separating (median delta = -0.010). A cross-validated logistic model restricted to the three out-of-rule variables reaches a median AUC of 0.789 across the nine corpora, against 0.924 for a model built on the rule

inputs. The rules therefore select accounts that differ on evidence the rules did not use, which is the informative part of the result.

The separation is nonetheless a cut on a continuum, not the isolation of a detached cluster. Bimodality coefficients exceed the 0.555 reference threshold for most variables in most corpora and a two-component Gaussian mixture beats a single component on BIC, but neither diagnostic is informative here: the distributions are degenerate, with the modal value holding a median of 89.3% of accounts for retweet share, 99.8% for intra-account repetition and 100% for coordinated-text share and verification. The apparent two-mode structure of retweet share is an artifact of granularity, since a median of 75.7% of accounts contribute exactly one mention and their retweet share is therefore exactly 0 or exactly 1 by construction, with a median of only 2.5% of accounts lying anywhere in between. Overlap confirms the picture: OVL medians run from 0.79 to 1.00 on seven of the nine variables, so outside retweet share and hashtags per mention the two groups occupy the same region of the feature space.

That control corpus reproduces this geometry instead of dissolving it, and on the variable that drives the split it reproduces it more sharply than the median case corpus: retweet share separates with delta = 0.646 and OVL = 0.36 in the control, against a median delta = 0.492 and OVL = 0.51 across the nine case corpora. Out-of-rule separation also persists there, with follower count at delta = -0.182 and coordinated-text share at delta = 0.264, both with $p < 0.001$, and with 32.2% of control Trolls attached to at least one coordinated text against 6.5% of control non-Trolls. Handle-defined Trolls in the control differ from matched peers with the same sign and comparable magnitude as in the case corpora (retweet share delta = 0.526, follower count delta = -0.317). The account-level category is thus a behavioral stratum, made of low-follower, retweet-only, single-mention accounts carrying a clear Stance, that exists in any large conversation on X, a football derby included. Its presence is not by itself evidence of an operation, and neither is its size: the control corpus reaches D1 = 34.1%, above three of the ten case corpora (6.5%, 23.9% and 29.4%), while the rest of its panel stays low (D2 = 10.0%, D3 = 6.3%, D4 = 0.0%).

The matched contrast was computed over the ten study corpora and over the first organic control corpus. The three remaining control corpora entered the study after that computation and are reported at panel level in Section 5.3.

Table S21. Matched within-corpus contrast between accounts classified as Trolls and an equally sized random sample of the accounts of the same corpus not so classified; the smaller of the two arms enters whole and the larger is subsampled. Panel A uses all Trolls; Panel B restricts the Troll arm to accounts whose label is decided by the handle rules (R1, R2). Statistics are medians over 500 matched draws per corpus; the case corpora are summarized by the median and range of the per-corpus statistic over nine corpora (Dominican Republic EV001 excluded for insufficient arm size). delta is Cliff's delta, OVL the overlapping coefficient, and k the number of corpora in which the contrast reaches $p < 0.05$ in at least 95% of draws.

Variable	Rule input	Nine case corpora: median Troll	median non-Troll	delta (range)	k
Panel A. All Trolls					
Mentions contributed	Yes (R3, R4, R5, R6)	1.0	1.0	0.005 (-0.428 to 0.139)	4
Retweet share	Yes (R4, R6)	1.00	0.00	0.492 (0.000 to 0.992)	7
Coordinated-text share	No	0.00	0.00	0.000 (0.000 to 0.693)	4
Intra-account repetition	Yes (R5)	0.00	0.00	0.000 (-0.035 to 0.015)	0
Aggressive-lexicon share	Yes (R3)	0.00	0.00	0.000 (-0.094 to 0.700)	5
Handle digit-pattern score	Yes (R1, R2)	0	0	0.181 (0.103 to 0.909)	9
Platform verification	Yes (R2, R3, R6)	0	0	0.000 (-0.262 to 0.000)	2
Followers	No	276	716	-0.102 (-0.345 to -0.012)	5
Hashtags per mention	No	0.00	0.62	0.062 (-0.749 to 0.785)	6
Panel B. Trolls labeled by the handle rules					
Retweet share	Yes (R4, R6)	1.00	0.00	0.241 (0.000 to 0.953)	6
Coordinated-text share	No	0.00	0.00	0.000 (0.000 to 0.685)	4
Followers	No	119	716	-0.324 (-0.515 to -0.195)	6
Hashtags per mention	No	0.00	0.62	0.053 (-0.525 to 0.582)	6

File S11. Stance classifier specification

Stance is assigned by counting literal pattern occurrences against a case-specific lexicon, with the decision rule stated in subsection 3.1. Each lexicon declares three pattern lists and one narrative dictionary. Attack patterns and defense patterns are the two lists the decision rule consumes. The neutral-marker list is declared for curation purposes and is reported below for completeness, but the decision rule does not consume it: a text that activates no attack and no defense pattern falls into "No stance" through the tie branch, whether or not it also matches a neutral marker. The narrative dictionary is a separate structure, consumed at the fourth stage of the workflow, that maps each mention to a dominant interpretive frame by first-match precedence over an ordered list of frames with an "Others / general" residual; its frames are reported in Table 7 and its pattern load in File S5.

Two distinct normalizations operate on the text and the difference between them is substantive. The identity normalization lowercases the

text and removes the leading retweet prefix, preserving diacritics, punctuation and emoji; the digest of the result is the text identifier the pipeline uses to group replications. The matching normalization additionally decomposes the string into Unicode NFD form and drops every combining mark, and is applied to the corpus text and to the lexicon patterns alike, so that a pattern written unaccented matches an accented occurrence. The separation answers two different requirements. Text identity must stay conservative, because it defines what counts as a replication of the same text and therefore feeds the coordination condition of Equation (1); pattern matching must stay tolerant, because accentuation in platform writing is unstable. The conservative choice costs nothing measurable here: the 8,377 unique texts of the ten study corpora remain 8,377 distinct texts under diacritic folding, so no replication cluster is fragmented by accent variation alone.

Table S22. Composition of the stance lexicon of each study case and its empirical yield. Coded stance is the share of the mentions of the corpus receiving a directional code, "Attacks" or "Defends", rather than "No stance". Unique texts is the number of decisions the classifier issues in that corpus, since stance is decided once per text and propagated to its mentions.

Case	Attack patterns	Defense patterns	Neutral markers	Unique texts	Mentions	Coded stance (%)
EC/EV001 Noboa–Uribe	67	48	5	176	9,708	85.2
EC/EV002 J. J. Neira	16	0	14	25	197	33.0
EC/EV003 I. Vélez	20	0	10	394	4,062	99.9
EC/EV004 C. A. Vera	14	8	3	226	1,031	85.5
EC/EV005 Banco Pichincha	23	13	4	5,477	32,381	70.1
DO/EV001 Yayo Sanz	27	8	4	31	31	100.0
DO/EV002 Gloria Reyes	20	13	6	764	919	37.9
MX/EV001 A. Armenta	32	10	4	943	5,765	100.0
MX/EV002 F. Noroña	21	10	4	115	130	100.0
MX/EV003 Clara Brugada	18	7	5	226	2,399	99.2
Ten cases	258	117	59	8,377	56,623	78.8

Three properties of the table deserve comment. Lexicon size does not track corpus size: the Banco Pichincha corpus, the largest at 32,381 mentions, is coded with 36 directional patterns, while the Noboa–Uribe corpus, at 9,708 mentions, requires 115, because the latter carries two mutually hostile camps and the former is dominated by a single direction of attack. Coverage does not track lexicon size either: it tracks how concentrated the conversation is on the target actor, which is why the two corpora with the lowest coverage are the two in which the conversation about the actor is embedded in a wider institutional debate. And two of the ten lexicons, EC/EV002 and EC/EV003, declare an empty defense list, so in them the rule degenerates to a binary decision between "Attacks" and "No stance" and can never return "Defends"; that is a property of those two lexicons and not of the classifier.

One limitation of provenance is recorded. Two of the ten lexicons, EC/EV002 and EC/EV003, were reconstructed by distinctive n-gram

extraction over the preserved classified mentions after the original pattern files were lost in a repository migration. For those two cases the coverage figure is partly self-referential, since the reconstructed patterns were derived from the labels they reproduce, and it should not be read as an independent measurement of lexicon quality. The remaining eight lexicons were curated from corpus inspection prior to classification.