Formation of a Multilevel System of Behavioral Indicators for Digital Analytics

Mykhailo Otmorskyi*, Andrii Roskladka

Department of Digital Economy and System Analysis, State University of Trade and Economics, Kyiv, Ukraine

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1311000074

Received: 04 November 2026; Accepted: 09 November 2026; Published: 24 November 2026

ABSTRACT

Behavioral dashboards may silently convert features of individual digital records into claims about communities, even when observation units, denominators and aggregation operators are unspecified. This study specifies a three-level behavioral-indicator framework and evaluates its arithmetic consistency and traceability in a reproducible controlled computational experiment. A corpus of 400 deliberately generated records was allocated to two simulated communities and four calendar-labelled weekly bins (3–24 August 2026); those dates are experimental indexing conventions, not dates of observed platform activity. Record-level ground-truth flags denote cooperative expressions, hostile expressions, explicit distrust and three-category stance. Fully specified equations produce group-level rates, within-group stance coexistence and separate between-group stance divergence. Every group–week result retains its denominator, scale and processing identifiers; percentile bootstrap intervals are calculated using 2,000 resamples per stratum. From the first to the fourth simulated week, cooperative-expression shares decline in both groups (0.80 to 0.70 and 0.20 to 0.10), whereas the record-weighted pooled share rises from 0.26 to 0.64 because group composition changes. Equal-community aggregation instead declines from 0.50 to 0.40. An exact decomposition of the pooled change allocates +0.48 to composition and −0.10 to within-group changes. A separate TF-IDF/logistic-regression check demonstrates execution of the text-processing stage on generated English templates; its held-out scores cannot be interpreted as validity on naturally occurring communication. The central contribution is an inspectable correspondence between record-level measurements, grouping decisions and macro summaries, not evidence about real communities or population behavior. The released generator, records, metadata, tables and calculation scripts permit independent reproduction of all experimental results. These distinctions establish the appropriate evidentiary boundary of the reported numerical findings.

Keywords: behavioral indicators; multilevel measurement; Simpson’s paradox; group composition; text classification; reproducible digital analytics

INTRODUCTION

Digital traces may contain timestamps, sources, textual messages and interactions. When record-level scores are transferred to community dashboards without explicit grouping keys and denominators, changes in record composition may be mistaken for changes in underlying behavior [8; 13]. The earlier version of this manuscript described a micro–meso–macro hierarchy but did not supply sufficient feature-to-index specifications and used a single small demonstration. The revised study addresses that methodological gap through newly defined observable proxies and transparent controlled tests. Its outputs must be understood as results of computational test cases, not empirical observations of online communities.

The objective is to specify and verify a reproducible record-to-community-to-system analytical workflow. The verification covers defined proxy variables and scales, record eligibility, group denominators, processing versions, uncertainty calculations, a separate between-group statistic, four consecutive synthetic weekly bins,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

and a controlled comparison between pooled and equal-community estimands. A separate generated-text classification task checks implementability of the micro text-feature pipeline. No population-level inference or field accuracy claim is tested.

LITERATURE REVIEW

Computational social science treats digital traces as informative but non-representative evidence of collective behavior [8; 13]. Measurement validity depends on the relation between a theoretical concept and its operational indicator [1]; the availability of a computational score does not itself validate trust, conflict or polarization. Composite-indicator methods require documented scales, weighting, uncertainty and sensitivity [10]. Polarization measures address the distribution of opinions, rather than serving as synonyms for negative sentiment [6]. Ecological inference can be misleading when group composition is unobserved [12]. Reproducible datasets and quality procedures also require explicit documentation and provenance [7; 14].

Prior work on contextual language representations, including BERT, Sentence-BERT and Ukrainian-language resources, motivates the separation of NLP feature extraction from measurement and aggregation [4; 5; 11; 15; 16]. The present experiment does not implement these models. Its only text-processing check is TF-IDF plus logistic regression on generated English templates (Supplementary Table S1), which confirms that the software pipeline can run. It does not assess sentiment or toxicity detectors, embeddings on real messages, or predictive performance in Ukrainian.

Table 1: Research strands and the measurement gap

Strand	Established capability	Remaining constraint
NLP and language representations [4; 5; 11]	Text features and classifiers	Domain transfer; annotation and construct validity
Measurement and composite indices [1; 10]	Scale, weights and transparent indices	Proxy interpretation and uncertainty
Polarization and ecological inference [6; 12]	Distributional and group-sensitive statistics	Choice of analytical grain
Data quality and provenance [7; 14]	Auditable transformation workflows	Versioned computation and reproducibility

Table 1 identifies the specific gap addressed by the study. Language-processing techniques provide record-level attributes, whereas measurement-validity research warns against interpreting such attributes as collective behavior without a defined construct [1; 12]. Composite-indicator and data-documentation literature further motivates explicit scales, group denominators and processing provenance [7; 10; 14]. Consequently, the proposed contribution is not a new NLP classifier, but a traceable link between observation levels and aggregation rules.

METHODOLOGY

Observation grain, eligibility and record-level variables

Each generated record i has record_id, community_id and week_start; the latter is a Monday date assigned by the test protocol, not a timestamp collected from a real platform. The deterministic generator sets binary reference labels c (cooperative expression), h (hostile expression), d (explicit distrust) and ordinal stance z. Binary flags may co-occur. These labels are controlled test inputs; they are not psychological assessments, independently annotated real messages or outputs of an externally validated detector. The community–week key specifies the observation grain used in subsequent aggregation.

$$x_i = (c_i, h_i, d_i, z_i),\ c_i, h_i, d_i \in \{0,\ 1\},\ z_i \in \{-1, 0, 1\} \qquad (1)$$

The quality gate Q_i is the conjunction of required-field, valid-week, text sufficiency and unique-ID checks. In the four-week controlled corpus, all 400 generated records pass the specified checks, as independently

logged in quality_report.json. This is a functional test, not a measure of classifier accuracy or population generalizability.

$$Q_i = q_{required,i} * q_{time,i} * q_{length,i} * q_{dedup,i} \quad (2)$$

For nonempty groups, eligible records and their denominator N are defined as follows. Empty strata are stored as missing values rather than zeros.

$$D_{g,t} = \{i: Q_i = 1, community_i = g, period_i = t\}, \quad N_{gt} = |D_{g,t}| \quad (3)$$

Operational indicator definitions and calibration

All binary features have fixed support {0,1}, and stance has fixed support {−1,0,+1}; no learned weights, estimated min–max bounds or undocumented calibration constants enter these definitions. Cooperation C and hostile-expression H are shares of flagged records. They are not respectively validated measures of prosocial cooperation and interpersonal conflict.

$$C_{gt} = \frac{\sum_{i \in D_{gt}} c_i}{N_{gt}}, \; H_{gt} = \frac{\sum_{i \in D_{gt}} h_i}{N_{gt}}, \quad (4)$$

Explicit-distrust share T is distinct from validated psychological trust. Mean stance μ retains direction and may be negative. For an individual group and period, coexistence of positive and negative stances is operationalized with 4 p+ p−. Neutral records remain in N; hence this score is suppressed when many messages are neutral. It is a bounded descriptive proxy rather than a validated distributional polarization index.

$$T_{gt} = \frac{\sum_{i \in D_{gt}} d_i}{N_{gt}}, \; \mu_{gt} = \frac{\sum_{i \in D_{gt}} z_i}{N_{gt}}, \quad (5)$$

$$P_{gt}^{within} = 4_{gt}^{+} p_{gt}^{+}, \; p_{gt}^{\pm} = \frac{\sum_{i \in D_{gt}} 1[z_i = \pm 1]}{N_{gt}} \quad (6)$$

For exactly two communities, D_between equals half the absolute distance between their mean stances and belongs to the macro grain; it cannot be attributed to a single meso row. Values of different indices must not be ranked as relative behavioral intensities even when their numerical support is the unit interval.

$$D_t^{between} = \frac{|\mu_{At} - \mu_{Bt}|}{2} \in [0,1], \quad (7)$$

Table 2. Fully specified indicators, parameters and aggregation grain

Indicator	Required inputs	Transformation and fixed range	Reporting grain
Cooperation C	c ∈ {0,1}; N>0	Mean(c), [0,1]; no weights	Group × week
Hostile expressions H	h ∈ {0,1}; N>0	Mean(h), [0,1]; no weights	Group × week
Explicit distrust T	d ∈ {0,1}; N>0	Mean(d), [0,1]; no weights	Group × week
Opposing-stance coexistence P	z ∈ {−1,0,+1}; N>0	4 × share(+) × share(−), [0,1]	Group × week
Between-group divergence D	Mean stance of A and B	|μ_A−μ_B| / 2, [0,1]	Two groups × week

Table 2 makes the operational meaning of each reported quantity explicit. Cooperation (C), hostile-expression share (H) and explicit-distrust share (T) are proportions of controlled binary labels; within-group stance coexistence (P) uses the positive and negative stance proportions, while between-group divergence (D) compares mean stances of two groups. The inputs and fixed ranges require no learned weights or estimated

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

calibration limits. Identical numerical ranges do not establish that these different proxies are socially interchangeable or validated behavioral constructs.

Alternative aggregation targets and uncertainty

Pooled cooperation is the mean for a uniformly sampled eligible record. Equal-community cooperation gives each represented community equal weight. For mean-based meso values, pooled aggregation is mathematically identical to the average of all records; that identity is not a novel estimator. The contribution is the ability to retain both estimands and distinguish them in outputs. An analyst chooses the estimand according to the research question.

$$C_t^{pooled} = \frac{\sum_g N_{gt} C_{gt}}{\sum_g N_{gt}}, C_t^{balanced} = \frac{1}{G_t} \sum_g C_{gt} \qquad (8)$$

The retention rate \(R\) represents the proportion of records that satisfy the established data quality criteria, whereas \(E\) denotes the absolute difference between a calculated estimate and its known reference value. For each group–week stratum, 95% percentile bootstrap confidence intervals are calculated using 2,000 resampling iterations with replacement. The resulting intervals quantify the variability of the estimated behavioral indicators under the specified resampling procedure. For the pairwise macro-level statistic, confidence intervals are obtained by independently resampling observations within each community and recalculating the corresponding aggregate measure. Table 6 presents the values of all four behavioral indicators and their associated 95% bootstrap confidence intervals across eight group–week strata.

$$R = \frac{N_{accepted}}{N_{input}}, E = |\hat{\theta} - \theta_{known}| \qquad (9)$$

$$CI_{95\%}(\hat{\theta}) = [Q_{0.025}(\hat{\theta}^*), Q_{0.975}(\hat{\theta}^*)] \qquad (10)$$

Traceable architecture and data model

The proposed logical architecture separates original generated records, micro labels, weekly group estimates, macro summaries and processing metadata. Outputs retain run_id, feature_version, calibration_version, grouping keys and the relevant denominator. Figure 3 uses the logical name fact_behavioral_indicators, not the legacy fact_sustainability identifier from the dissertation prototype. It is a design schema for the present experiment, not a claim that the original Power BI deployment executed the new formulas. community_id identifies groups at micro and meso grain; macro summaries refer to the set of compared communities. The distrust flag is a record-level test label contributing to a group-level share, not an observed population-level psychological attribute.

Figure 1. Record-to-macro workflow with processing metadata

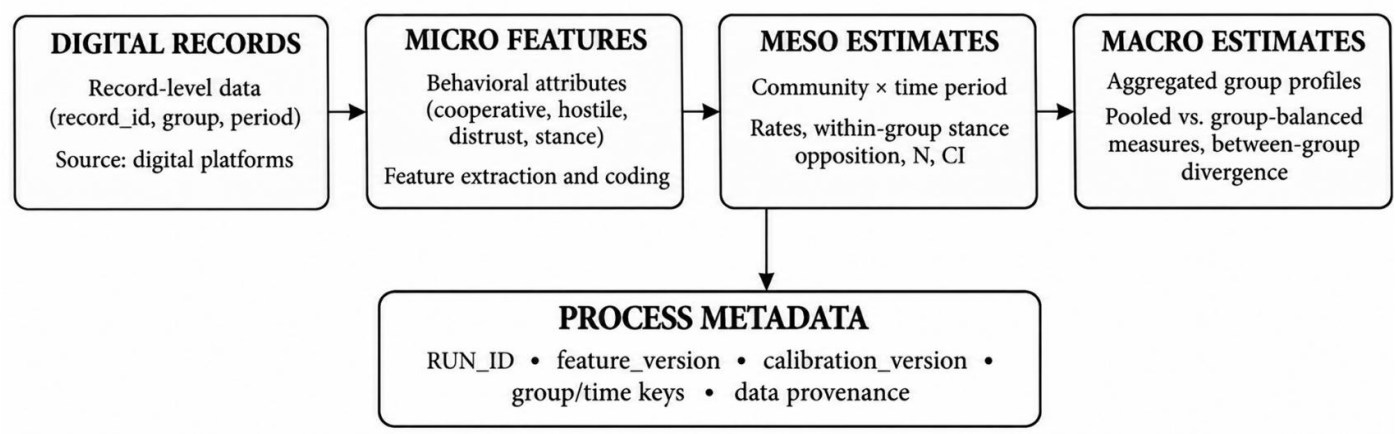


Figure 1 illustrates the change in observation unit from an eligible digital record to a record-level feature vector, a community-by-week estimate, and a cross-community summary. The bottom metadata component

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

connects the results to RUN_ID, feature and calibration versions, grouping keys and provenance information. The arrows specify data transformations and aggregation, not causal effects; preserving those keys makes it possible to distinguish a change in group composition from a change within groups.

Figure 2. Traceable preparation, label representation and analytical outputs

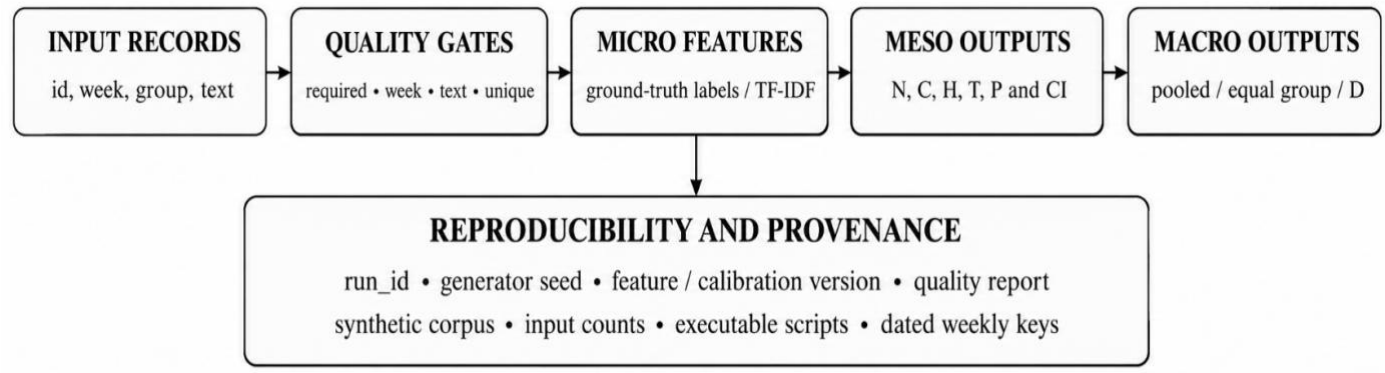


Figure 2 separates eligibility checks and feature representation from the calculation of meso and macro results. The generated reference labels feed the indicator definitions directly, while the ancillary TF-IDF/logistic-regression check is a separate software-integration procedure. Thus the main indicator values in Tables 3 and 4 are not attributed to a trained sentiment, toxicity or neural-embedding model. Retaining intermediary outputs allows each reported aggregate to be recomputed from the same records.

Figure 3. Logical data model for multilevel behavioral measurements

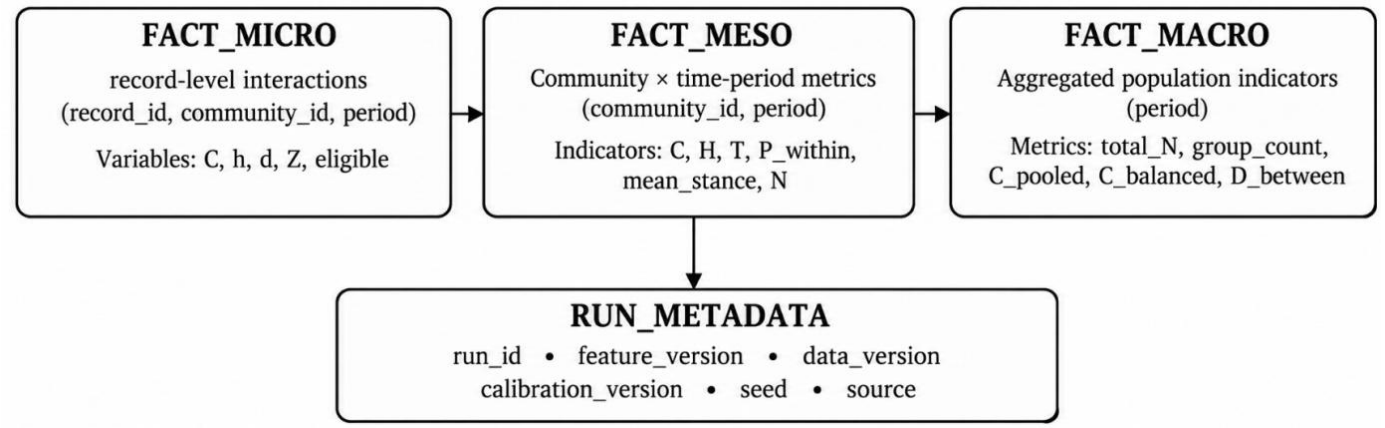


Figure 3 specifies the relational links needed for reproducibility: a record identifier at micro grain, community and week keys at meso grain, and period-indexed aggregation with the associated processing metadata. The name fact_behavioral_indicators designates the revised logical representation and replaces the potentially misleading sustainability label. This drawing is not a screenshot of the earlier Power BI deployment. A field such as community_id identifies groups in meso outputs; a binary distrust label is retained at record grain and must be aggregated explicitly before a macro interpretation is attempted.

Controlled synthetic corpus and software environment

The experimental corpus comprises 400 fully generated, individually identifiable test records, arranged in four synthetic seven-day bins labelled 2026-08-03, 2026-08-10, 2026-08-17 and 2026-08-24. These four labels facilitate longitudinal aggregation tests but do not represent four weeks of independently observed platform traffic. For each bin, the generator assigns records to simulated communities A and B, varies group composition and explicitly determines the reference proportion of cooperative messages. It assigns binary flags and ordinal stance before composing English messages from a small label-dependent template vocabulary. Because the reference labels are known by construction, the corpus can test exact arithmetic, denominator preservation and aggregation sensitivity. The full generator and record-level CSV are included

with the supplementary materials; the generation seed and software versions are logged. No real authors, social-media accounts or external Ukrainian-language posts were sampled.

The ancillary software-integration check applies scikit-learn TfidfVectorizer (word unigrams and bigrams; min_df=2; sublinear_tf=True) followed by four LogisticRegression classifiers to generated English template sentences. A 75%/25% train/test split uses random_state=20261113, and the TF-IDF vocabulary is fit only on the training partition. The label-dependent generator repeats lexical patterns across the split, creating a substantial shortcut-learning risk. Therefore, these scores test pipeline execution only; they are not estimates of natural-language model performance, classifier generalization, or construct validity. All numerical aggregation results in Tables 3–6 use reference labels directly and do not depend on the classifiers. No sentiment-specific or toxicity-specific detector, BERT/Sentence-BERT embedding model, or Ukrainian-language model was run in this experiment. The auxiliary classification diagnostics are provided as Supplementary Table S1 rather than as primary research evidence.

RESULTS

Weekly group-level indicators and traceability

The complete experimental code generated the records, calculated all proportions, retained N and the group–week keys, and materialized separate meso and macro CSVs. Each output was computed from the corresponding actual rows. For 2026-08-03, community A contains eight cooperative, two hostile, three distrust-labelled messages, six supportive stances and four opposed stances among ten messages. Its opposing-stance coexistence equals 0.96 and the between-community divergence with B equals 0.25:

$$P_{A,T1}^{within} = 4 * \frac{6}{10} * \frac{4}{10} = 0.96 \quad D_1^{between} = \frac{|0.20-(-0.30)|}{2} = 0.25 \qquad (11)$$

Table 3. Four-week meso outputs computed from released synthetic records

Week starting	Community	N	C	H	T	P within	Mean stance
8/3/2026	A	10	0.8	0.2	0.3	0.96	0.2
8/3/2026	B	90	0.2	0.4	0.1	0.4	-0.3
8/10/2026	A	30	0.767	0.233	0.2	0.8	0.267
8/10/2026	B	70	0.171	0.4	0.1	0.48	-0.4
8/17/2026	A	70	0.729	0.3	0.2	0.72	0.3
8/17/2026	B	30	0.133	0.4	0.1	0.48	-0.4
8/24/2026	A	90	0.7	0.3	0.2	0.72	0.3
8/24/2026	B	10	0.1	0.5	0.2	0.32	-0.7

Table 3 reports all eight community–week strata, rather than one pooled demonstration value. The N column gives the exact denominator used to calculate C, H, T and within-group P for each simulated group and calendar-labelled bin. Mean stance is displayed separately because it serves as an input to the between-group macro quantity. The four dates identify bins imposed by the experimental protocol; they are not dates of real platform observations. Row-level checks can be reproduced using the supplied CSV and calculation script.

The group-level outputs report N and the four defined observable proxies for each community and synthetic weekly bin. Mean stance and within-group coexistence are retained separately from between-group divergence, preventing a meso value from being mislabeled as a macro polarization measure. All displayed values are recomputed from the accompanying synthetic record-level flags and the published equations. The earlier four-index histogram and its undocumented historical values are omitted because they cannot be numerically reconstructed from the new model.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

Macro results, baseline and composition decomposition

Table 4. Weekly macro results and composition-sensitive baseline

Week starting	N	Pooled C	Equal-community C	Between-group D	N(A)	N(B)
8/3/2026	100	0.26	0.5	0.25	10	90
8/10/2026	100	0.35	0.469	0.333	30	70
8/17/2026	100	0.55	0.431	0.35	70	30
8/24/2026	100	0.64	0.4	0.5	90	10

Table 4 distinguishes the record-weighted pooled estimate from the equal-community estimate for each of four weekly bins. The first answers a record-level population question, whereas the second assigns equal analytical importance to each simulated community. For the controlled composition shift, pooled cooperation rises from 0.260 to 0.640 although each group-specific cooperation share decreases; the equal-community result falls from 0.500 to 0.400. The N(A) and N(B) columns show why these trends diverge. Between-community stance divergence is a separate macro quantity and is not a weighted average of meso polarization.

In each community cooperative-expression shares decline from the first to the last week (A: 0.800 → 0.700; B: 0.200 → 0.100). Nevertheless, the record-weighted total rises from 0.260 to 0.640 as community A's record share grows from 10% to 90%. By contrast, the equal-community estimate falls from 0.500 to 0.400. The first and last pooled means can be checked manually:

$$C_1^{pooled} = \frac{10(0.80)+90(0.20)}{100} = 0.26,\ C_{T1}^{balanced} = \frac{0.80+0.20}{2} = 0.50 \qquad (12)$$

The exact two-period symmetric decomposition of the pooled change assigns +0.480 to the change in group composition and −0.100 to within-group changes, resulting in +0.380 overall. This decomposition explains the counterdirectional trend without claiming that the equal-community statistic is universally preferable.

$$\Delta C^{pooled} = 0.48_{composition} - 0.10_{within} = 0.38 \qquad (13)$$

Figure 4. Four-week community and macro cooperation trends under composition change

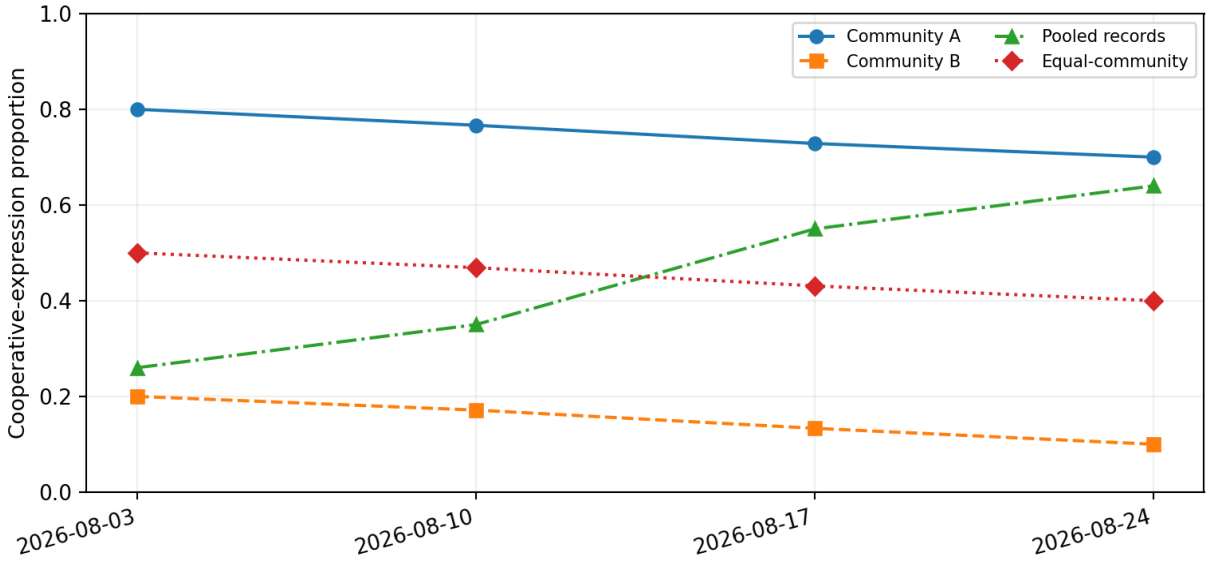


Figure 4 displays the four controlled weekly observations of cooperation shares for communities A and B alongside the record-weighted pooled estimate. Within each community the trajectory is downward, yet the pooled trajectory increases because the relative number of records contributed by the two groups changes. The

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue XI November 2026

graphic visualizes a composition-induced reversal; it does not represent an observed trend in real online communities.

Figure 5. Four-week pooled versus equal-community estimates

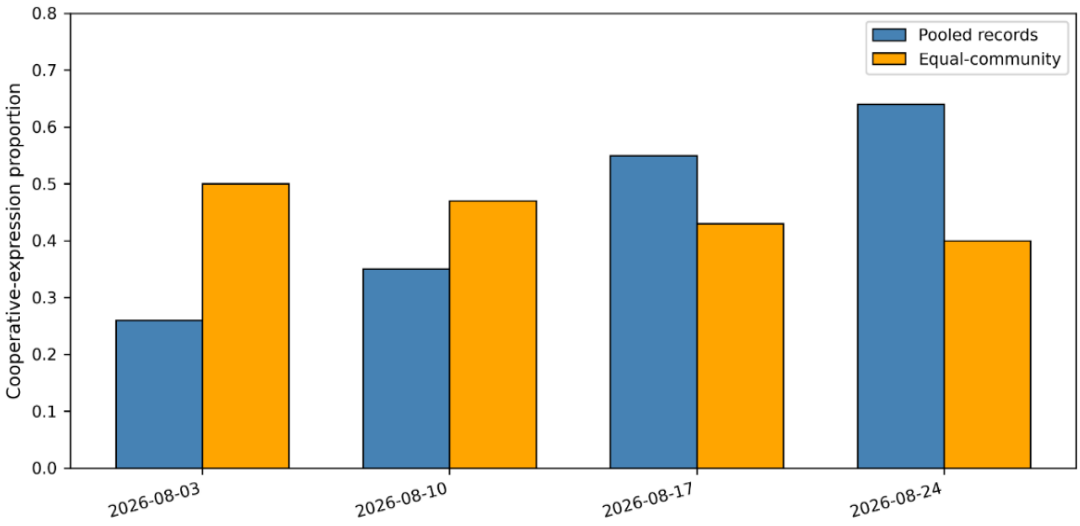
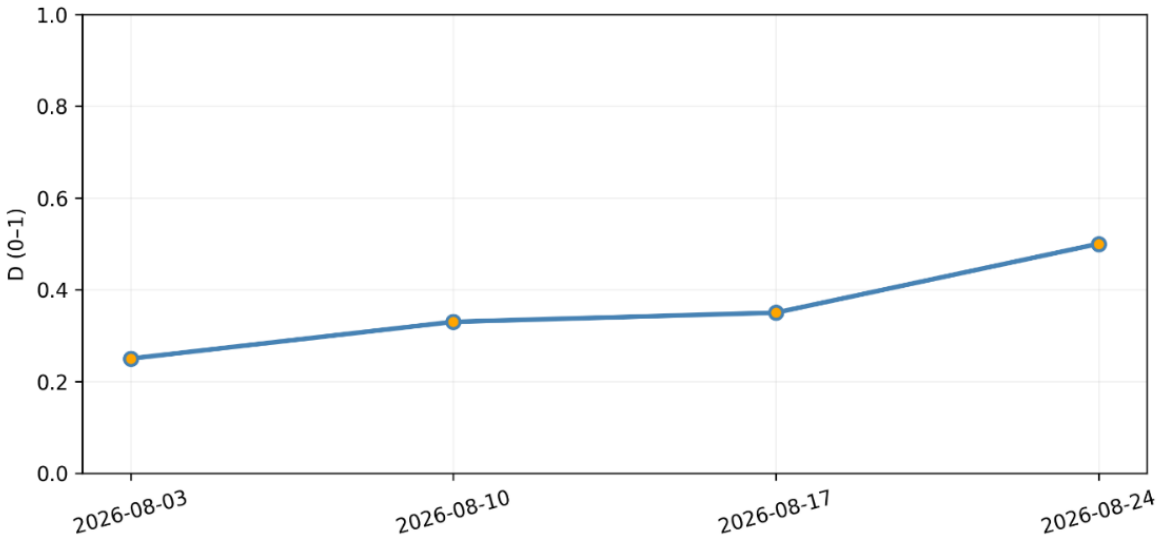


Figure 5 isolates the difference between pooled and equal-community aggregation. The pooled curve depends on the shifting numbers of records in A and B, while the equal-community curve reflects the average of their group-specific shares under a constant two-group weighting rule. The difference is a change in the estimand, not proof that equal-community weighting is always superior. Denominators for every week are provided in Tables 3 and 4.

Figure 6. Between-community stance divergence across four weeks



Source: Authors' computation of between-community mean-stance divergence from data/meso_weekly.csv; Equation (7).

Figure 6 illustrates the separately defined between-community divergence statistic D across the four simulated weekly bins. Unlike within-group stance coexistence P in Table 3, D is calculated from two community mean-stance values and therefore belongs to macro grain. Its time profile can be reproduced from the displayed mean stances; the chart does not establish a validated sociological measure of polarization.

Complete uncertainty estimates and volume sensitivity

Table 5. Volume invariance under fixed label proportions

N	Cooperation C	Hostile H	Distrust T	Stance coexistence P
100	0.4	0.1	0.1	0.25
500	0.4	0.1	0.1	0.25
1000	0.4	0.1	0.1	0.25

Table 5 holds the underlying reference-label proportions fixed while varying the number of generated records. The unchanged values are therefore a positive-control check of arithmetic consistency, not evidence that the measurement is invariant to changes in real communication behavior or to sampling bias. This test complements, rather than replaces, the composition-shift analysis in Table 4.

The values in Table 5 remain invariant by construction: each generated sample has the same target label proportions. The test therefore verifies arithmetic under changed sample volume; it does not establish sampling invariance, resistance to distribution drift or model accuracy in observed communication.

Table 6. 95% percentile bootstrap intervals for every meso indicator

Week	Group	N	Indicator	Estimate	CI lower	CI upper
8/3/2026	A	10	C	0.8	0.5	1
8/3/2026	A	10	H	0.2	0	0.5
8/3/2026	A	10	T	0.3	0	0.6
8/3/2026	A	10	P_within	0.96	0.36	1
8/3/2026	B	90	C	0.2	0.122	0.289
8/3/2026	B	90	H	0.4	0.3	0.5
8/3/2026	B	90	T	0.1	0.044	0.167
8/3/2026	B	90	P_within	0.4	0.255	0.533
8/10/2026	A	30	C	0.767	0.6	0.9
8/10/2026	A	30	H	0.233	0.1	0.4
8/10/2026	A	30	T	0.2	0.067	0.367
8/10/2026	A	30	P_within	0.8	0.507	0.982
8/10/2026	B	70	C	0.171	0.086	0.257
8/10/2026	B	70	H	0.4	0.286	0.514
8/10/2026	B	70	T	0.1	0.029	0.171
8/10/2026	B	70	P_within	0.48	0.294	0.651
8/17/2026	A	70	C	0.729	0.629	0.829
8/17/2026	A	70	H	0.3	0.2	0.414
8/17/2026	A	70	T	0.2	0.1	0.286
8/17/2026	A	70	P_within	0.72	0.539	0.869
8/17/2026	B	30	C	0.133	0.033	0.267
8/17/2026	B	30	H	0.4	0.233	0.567
8/17/2026	B	30	T	0.1	0	0.2
8/17/2026	B	30	P_within	0.48	0.178	0.733
8/24/2026	A	90	C	0.7	0.611	0.8
8/24/2026	A	90	H	0.3	0.211	0.4
8/24/2026	A	90	T	0.2	0.122	0.289
8/24/2026	A	90	P_within	0.72	0.56	0.847
8/24/2026	B	10	C	0.1	0	0.3
8/24/2026	B	10	H	0.5	0.2	0.8

8/24/2026	B	10	T	0.2	0	0.5
8/24/2026	B	10	P_within	0.32	0	0.84

Table 6 presents uncertainty intervals for every combination of the four meso indicators and eight simulated group–week strata. Each entry reports the point estimate, sample size and empirical 2.5th and 97.5th percentiles of the bootstrap distribution. The limits describe resampling variability within the prescribed synthetic design; they must not be interpreted as confidence intervals for unobserved real communities. Macro-level bootstrap outputs are supplied separately in data/bootstrap_macro.csv.

Table 6 reports percentile-bootstrap intervals for cooperation, hostile-expression share, explicit-distrust share and within-group stance coexistence in each of eight group–week strata. The supplementary bootstrap_macro.csv similarly contains intervals for the four between-group divergence results. Since all samples are controlled generated records, these intervals are conditional computational descriptors only. Analysis of real networked messages would require a sampling design that accounts for repeated authors, dependence and collection bias.

Ancillary text-processing software check

The auxiliary TF-IDF/logistic-regression scores are reported in Supplementary Table S1, not in the main experimental tables.

This optional test checks feature-vector creation, model fitting and prediction on label-dependent English templates. It is not a comparison of sentiment, toxicity or neural-embedding approaches, and its near-perfect results are not external NLP performance evidence. The primary analysis evaluates aggregation with known reference labels; it neither estimates classification accuracy on naturally occurring messages nor depends on the optional classifier outputs.

DISCUSSION

The demonstrated result is a traceable computation and a composition-induced reversal in pooled measurements under an explicit synthetic test design. The mathematical definitions specify observable features but do not prove construct validity: within-group P captures joint occurrence of opposed labels, and D_between captures separation of two groups' mean stance scores. Neither is a general validated measure of societal polarization. Likewise, hostile and distrust labels represent assigned linguistic categories, not attributes of people. Transparent formulas and denominators remove computational ambiguity without solving selection, annotation and measurement-validity problems [1; 6; 10].

Four dated synthetic bins and macro results directly address the request for an explicit time-indexed demonstration, but they cannot establish real weekly trends. Pooled and equal-community summaries estimate different targets; equal-group weighting is not a universal correction for Simpson's paradox. The decomposition in Equation (13) makes the contributions of group representation and within-group change auditable. Uncertainty values are conditional on the generated sample and resampling scheme, not estimates of uncertainty in a sampled human population.

The auxiliary TF-IDF/logistic-regression check is documented in Supplementary Table S1 solely as a software smoke test. Because the generated phrases and target labels are structurally linked, a random record-level holdout does not constitute independent domain evaluation. The experiment does not train or evaluate sentiment-specific, toxicity-specific, BERT/Sentence-BERT or Ukrainian-language models. These were described in the literature because they can supply record features in future applications, not because they were executed here. The present evidence supports arithmetic consistency, transparent grouping, recorded denominators, and the behavior of alternative aggregation targets under controlled composition changes. No claims are made about accuracy on observed communications or validity of social constructs.

CONCLUSION

This study presents fully operationalized micro-, meso- and macro-level calculations and demonstrates their execution on 400 deterministic synthetic records in two simulated communities across four weekly indexing bins. Each of four group-level proxies has fixed inputs, scale and operator, while a separate pairwise statistic captures between-group stance divergence. Reproducible scripts, denominators, all 32 group-level estimates with conditional bootstrap intervals, and four macro summaries are provided.

Across the controlled bins, cooperative-expression shares fall in both simulated groups while the pooled result rises from 0.260 to 0.640 because the groups contribute different numbers of records. An exact decomposition separates the +0.480 composition contribution from the −0.100 within-group contribution; equal-community weighting yields a different trajectory. A simple TF-IDF/logistic-regression test establishes software executability on label-revealing generated sentences, not real-world recognition quality. The findings support algorithmic reproducibility and transparent interpretation of aggregation targets, not conclusions about actual digital communities or predictive superiority.

DATA AND CODE AVAILABILITY

The computational experiments are supported by a reproducibility package containing the Python implementation, synthetic record-level data, intermediate analytical outputs, statistical estimates, and execution metadata. The main script (code/reproduce.py) implements the controlled data-generation procedure, record-level indicator calculations, micro-, meso-, and macro-level aggregation, and uncertainty estimation.

The accompanying datasets include 400 generated records distributed across two simulated communities and four weekly observation bins. Separate CSV files contain group-level behavioral indicators, macro-level aggregation results, bootstrap confidence intervals, and volume-sensitivity calculations. The experimental protocol documents the data-generation parameters, processing configuration, and random seed, while the quality report records the results of the specified data-validation checks.

The supplementary materials also include the results of the TF-IDF and logistic-regression software verification experiment. These results are provided separately from the main aggregation experiments.

All reported computational results can be independently examined using the accompanying data files and calculation procedures. The complete reproducibility package is provided as supplementary material with the revised manuscript.

REFERENCES

1. Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. American Political Science Review, 95(3), 529–546. https://www.cambridge.org/core/journals/american-political-science-review/article/abs/measurement-validity-a-shared-standard-for-qualitative-and-quantitative-research/91C7A9800DB26A76EBBABC5889A50C8B
2. Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. CIDR. https://www.vldb.org/cidrdb/2021/
3. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/
4. Chaplynskyi, D. (2023). Introducing UberText 2.0: A corpus of modern Ukrainian at scale. Proceedings of UNLP, 1–10. https://doi.org/10.18653/v1/2023.unlp-1.1
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423
6. Esteban, J.-M., & Ray, D. (1994). On the measurement of polarization. Econometrica, 62(4), 819–851. https://ideas.repec.org/a/ecm/emetrp/v62y1994i4p819-51.html

7. Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723
8. Lazer, D. M. J., Pentland, A., Watts, D. J., et al. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170
9. Newman, M. E. J. (2003). The structure and function of complex networks. SIAM Review, 45(2), 167–256. https://doi.org/10.1137/S003614450342480
10. OECD / European Union / EC-JRC. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing. https://doi.org/10.1787/9789264043466-en
11. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. EMNLP-IJCNLP, 3982–3992. https://doi.org/10.18653/v1/D19-1410
12. Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15(3), 351–357. https://doi.org/10.2307/2087176
13. Salganik, M. J. (2018). Bit by bit: Social research in the digital age. Princeton University Press. https://www.bitbybitbook.com/
14. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), 1781–1794. https://doi.org/10.14778/3229863.3229867
15. Shynkarov, Y., Solopova, V., & Schmitt, V. (2025). Improving sentiment analysis for Ukrainian social media code-switching data. Proceedings of UNLP, 179–193. https://doi.org/10.18653/v1/2025.unlp-1.18
16. Vysotska, V., Chyrun, L., & Bychkov, I. (2026). Information technology for detecting suicidal and depressive intentions in social networks. Visnyk VPI, 3, 17–26. https://doi.org/10.31649/1997-9266-2026-186-3-17-26