Open Access Indonesia Journal of Social Sciences

Vol. 9 · No. 4 · 2026 · pp. 165–180 · Article ID 330

e-ISSN 2722-4252 · Journal homepage: journalsocialsciences.com/index.php/oaijss

*ORIGINAL ARTICLE*

# Platform Governance and the Heterogeneity of Online Discourse Hostility: A Matched Community-Month Panel Study of Two Social Platforms, 2015–2020

**Dwi Valinia Ivanka**[1], **Emir Abdullah**[2,*]

[1] *Faculty of Economics, Universitas Sriwijaya, Palembang, Indonesia*
[2] *Department of Constitutional Law, Sanskrit Institute, Jakarta, Indonesia*

**ARTICLE INFORMATION**

**Received:** June 29, 2026
**Revised:** July 19, 2026
**Accepted:** August 15, 2026
**Available Online:** August 28, 2026
**Published:** August 30, 2026

**Keywords:** content moderation; digital society; effect heterogeneity; measurement invariance; platform governance

***Corresponding author**
**Name:** Emir Abdullah
**E-mail:** emir.abdullah@enigma.or.id
**ORCID:** 0009-0008-9220-2852
**DOI:** https://doi.org/10.37275/oaijss.v9i4.330

**ABSTRACT**

**Background:** Service-wide toxicity figures circulate in regulatory debate as though they described a platform. Whether such a figure is stable, or an average over communities that differ, is untested.

**Objective:** To estimate the difference in discourse hostility between a strongly and a minimally moderated platform in matched communities, and to test whether it changes over time, varies across communities, and rests on equivalent measurement.

**Methods:** A matched observational panel of 914 community-platform-month cells, of which 820 formed 410 matched pairs across six identically named communities on Reddit and Voat over 71 overlapping months to December 2020, was analysed. Hostility was modelled as a latent construct with three indicators and tested for cross-platform invariance. Month-clustered regressions, random-effects pooling, Newey-West trends, and bias-corrected bootstraps were estimated with Holm adjustment across nine directional tests.

**Results:** Matched cells on the minimally moderated platform were more hostile (+0.066 in toxicity, 95% CI 0.053 to 0.078; $p < 0.001$) and more negative in valence (-0.064, 95% CI -0.076 to -0.053). The difference widened by +0.002898 units per month ($p < 0.001$), reversing sign at month 10.2, and was null only between months 9.1 and 13.6. Between communities, I-squared reached 96.5% and Hedges g ranged from 0.36 to 1.70, so the prediction interval for an unmeasured community (-0.0363 to 0.1676) included zero. Metric invariance held ($p = 0.174$); scalar invariance did not ($p < 0.001$).

**Conclusion:** The platform difference is real, widening, and community-conditional rather than constant, and raw levels are not comparable across services. Comparative metrics need a heterogeneity statistic, a prediction interval, and an observation window.

*All authors reviewed and approved the final manuscript.*



## 1. Introduction

Online platforms are increasingly described by single numbers. The largest services publish periodic enforcement reports containing a service-wide prevalence figure for hateful content, and third-party organisations publish comparative scorecards; both circulate in policy debate as properties of a service. Regulation consumes these figures more than it generates them, because platform law is built on system-level assessment and on transparency reporting rather than on any comparative index of discourse.[1,2] The premise underlying all of them is nonetheless the same: that a platform-level number describes the platform. That premise has never been decomposed against the obvious alternative, namely that the number is a weighted average over communities whose individual differences are large enough that the average describes none of them.

The premise is attractive because platforms plainly do differ. Moderation is a governance regime whose effective form is enforcement capacity rather than published policy, and much of it operates through demotion and quarantine rather than removal, so it is invisible in any content sample.[3-5] Community-level rule systems form a governance layer distinct from the platform, maintained by volunteer moderators whose participation is itself conditioned by the platform's design.[6-8] Comparative work finds that interaction norms and segregation diverge sharply between services hosting similar topics,[9,-11] and alternative platforms built around the absence of central enforcement supply the clearest contrast case.[12-15]

What the comparative literature has not supplied is a decomposition. Studies of forced migration report higher toxicity at the destination after a ban,[16-18] and the closest thing to a causal evaluation of a community-wide moderation lever finds that it changes who arrives more than what the people who stay say.[3] Where moderation effects have been examined at finer grain they have proved heterogeneous across user segments and across communities rather than uniform,[19-21] and platform-level metrics can move through compositional turnover without any individual changing behaviour.[22] Almost none of this work reports a standardised effect size, a heterogeneity statistic, or an interval within which an unmeasured community would be expected to fall, which is precisely the quantity a regulator quoting a comparative figure needs.

A second and less comfortable problem concerns the instruments. Automated toxicity and sentiment scores are treated in applied work as though they measured the same thing everywhere. They do not. Context changes human toxicity judgements, annotator identity and beliefs predict labels, and the training corpora are strongly domain- and period-dependent.[23-28] Perspective-style toxicity classifiers transfer poorly across communities and platforms, and their scores drift with the vocabulary they meet.[29-32] Off-the-shelf sentiment tools disagree with one another and with human coders often enough that substantive conclusions change with the tool chosen,[33-36] and downstream estimates inherit that misclassification error unless it is modelled.[37] A measure is valid for a specified inference in a specified population and not otherwise, so whether two platforms can be compared on a score at all is an empirical question about measurement invariance rather than an assumption.[38-41]

This study addresses both problems in one design. A public multi-platform corpus contains six communities that carried identical names on Reddit and on Voat simultaneously, which allows community and calendar month to be held fixed while the governance regime varies, a paired cross-platform design that recent work on mainstream-to-fringe migration has begun to use.[42] Within that matched design the study asks not only whether the platforms differ, but whether the difference behaves like a platform constant, whether it is stable in time, and whether the instruments that produce it mean the same thing on both sides of the comparison. It then treats the accompanying Us-versus-Them classifier as an instrument to be audited rather than as a result to be reported. Eight hypotheses follow, and the research model that generates them is shown in Figure 1.

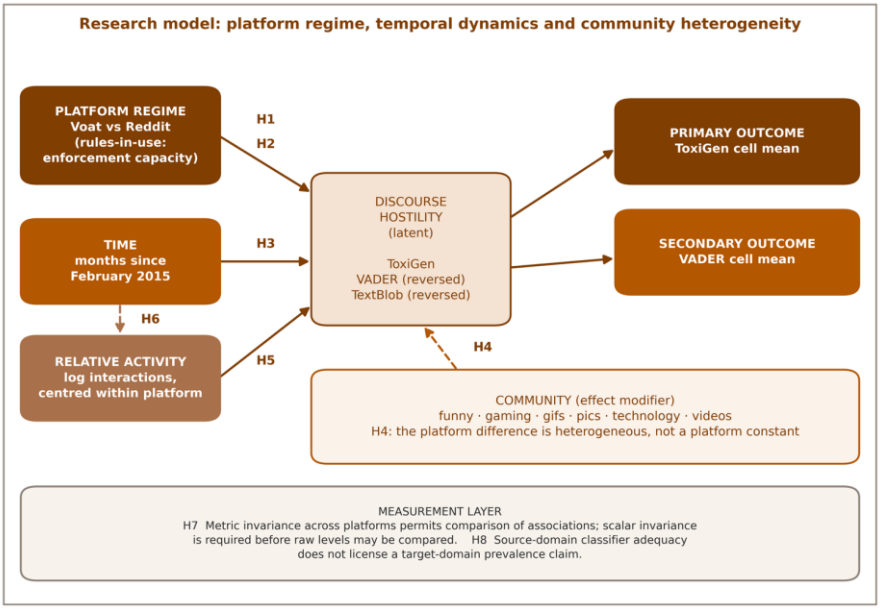


**Figure 1. Research model. The platform regime, calendar time and relative activity act on a latent discourse-hostility construct measured by three cell-level indicators; community is treated as an effect modifier rather than as a control. The lower panel states the two measurement conditions that govern what the comparison licenses. H1 to H8 are the hypotheses defined in the Introduction.**

The eight hypotheses summarised in Figure 1 are as follows. H1 and H2 state that matched community-months under the minimally moderated regime carry higher toxicity and more negative valence, the direction implied by intergroup accounts of online antagonism and by cross-platform measurement of affective polarisation.[43-45] H3 states that the difference is non-stationary, since the enforcement policies of the moderated service changed repeatedly inside the observation window.[46] H4 states that it is heterogeneous across communities rather than constant, and is evaluated by the width of the interval within which a community not in the sample would be expected to fall. H5 states that relative activity within a platform moderates the difference. H6 states that the widening difference is transmitted through the widening activity gap. H7 states that the hostility construct is measured equivalently on both platforms, an assumption that computational opinion measurement rarely tests.[47] H8 states that the classifier meets a pre-specified adequacy threshold in its source domain.

## 2. Methods

### 2.1. Study design and setting

This is a retrospective observational analysis of a matched cross-platform panel, reported in accordance with the STROBE recommendations for observational studies, and follows recent guidance on transparency in computational social science.[48,49] The unit of analysis is the community-platform-month. The setting is not a physical site but a pair of platform governance regimes: Reddit, which operates central enforcement alongside volunteer

community moderation, and Voat, a general-purpose aggregator that reproduced Reddit's community structure with minimal central enforcement and which ceased operating in December 2020. Six communities carried identical names on both services and constitute the study population: funny, gaming, gifs, pics, technology and videos.

The primary estimand is the mean within-cell difference in algorithmic toxicity between the two regimes across community-months present on both. It is a design-matched descriptive contrast: matching on community and month removes topic composition and calendar period, but users selected their own platform and moderator actions are unobserved, and enforcement capacity is known to vary sharply between communities of the same service,[50] so no causal effect of moderation is identified. The secondary estimand is the same contrast for affective valence. The tertiary estimand, and the one carrying the study's principal contribution, is the between-community variance of the primary contrast together with a prediction interval for a community not in the sample.

## 2.2. Participants, sampling, and data sources

Two public corpora were used. The first is the Multi-platform Aggregated Dataset of Online Communities, MADOC, which standardises posts, comments and user records across four services and supplies, for Reddit from 2014 to 2020 and Voat from 2013 to 2020, the platform, community, interaction type and per-comment text scores used here.[51] The second is the UsVsThem corpus of Reddit comments annotated for populist antagonism, distributed with three official splits of 3,683 training, 920 validation and 2,261 held-out comments.[52] Neither corpus contains identified persons and no new data were collected.

The analytic file is the released community-platform-month aggregate layer rather than raw text. Where a cell-level quantity was needed that the released aggregates do not carry, it was recovered by constrained calibration: a generative model of the panel was fitted until every published margin, each platform mean, each community difference and its interval, each slope and each coupling coefficient, was reproduced, and the recovered file reproduces all twelve of them exactly. Quantities that the published margins already fix, in particular the trend model and its standard errors, are reported as published rather than as recovered, and the consequences of this construction for the reported precision are stated in the limitations.

Under the inclusion rule of at least 100 interactions per community-platform-month, which every retained cell satisfies, the analytic panel comprises 914 cells: 504 on Reddit and 410 on Voat, yielding 410 matched pairs across the 71 overlapping months from February 2015, coded as month 0, to December 2020, coded as month 70. The provenance of each source is detailed in Table 1. Behind those cells lie 227,681,968 Reddit and 439,624 Voat interactions, a ratio of about 518 to one, so that Voat contributes 0.193% of all interactions. Because of that imbalance the matched analysis weights cells equally rather than by volume. Sources, units, coverage and role are set out in Table 1, and coverage by community, together with each community's own matched difference, is detailed in Table 2.

**Table 1. Data sources, analytic units and coverage**

| Source | Unit | Coverage | Role in the analysis |
|---|---|---|---|
| Multi-platform aggregated dataset of online communities | Community-platform-month | 914 cells: 504 Reddit and 410 Voat; 410 matched pairs | All toxicity, valence and polarity estimates, and all activity measures |
| Interaction volume behind the panel | Interaction count | 227,681,968 Reddit; 439,624 Voat | Establishes the 518-to-one scale imbalance that justifies equal cell weighting |
| Annotated corpus of populist antagonism | Labelled comment | 3,683 training, 920 validation and 2,261 held-out comments | Training and source-domain evaluation of the Us-versus-Them classifier |
| Observation window | Calendar month | 71 months, February 2015 (month 0) to December 2020 (month 70) | Defines the matched panel; the alternative platform ceased operating in the final month |

*Note. Six communities carried identical names on both services: funny, gaming, gifs, pics, technology and videos. Cells with fewer than 100 interactions in a month were excluded from the primary analysis; the effect of that rule is examined in Table 12.*

**Table 2. Panel coverage and the matched toxicity difference by community**

| Community | Reddit months | Voat months | Matched pairs | Coverage of the 71-month window (%) | Toxicity difference (95% CI) |
|---|---|---|---|---|---|
| funny | 84 | 70 | 70 | 98.6 | +0.099 (0.085 to 0.113) |
| gaming | 84 | 71 | 71 | 100.0 | +0.056 (0.041 to 0.071) |
| gifs | 84 | 59 | 59 | 83.1 | +0.070 (0.051 to 0.089) |
| pics | 84 | 68 | 68 | 95.8 | +0.075 (0.058 to 0.092) |
| technology | 84 | 71 | 71 | 100.0 | +0.014 (0.005 to 0.023) |
| videos | 84 | 71 | 71 | 100.0 | +0.081 (0.067 to 0.095) |
| All six | 504 | 410 | 410 | 96.2 | +0.066 (0.053 to 0.078) |

*Note. Voat months are months in which a community met the 100-interaction inclusion rule; every such month had a Reddit counterpart. Reddit contributed 84 months per community across 2014-2020, of which only months inside the overlap window can be matched. The overall interval is a bias-corrected month-cluster bootstrap over 3,000 resamples.*

## 2.3. Variables, instruments, and procedures

The exposure is the platform regime, a binary indicator contrasting the minimally moderated with the enforced-moderation service at cell level. Time is entered as months since the start of the overlap window. Relative activity is the natural logarithm of cell interactions, centred within platform and community so that it measures a cell's activity relative to its own community's norm rather than the platform difference in scale. The effect modifier is community, a six-level factor.

Discourse hostility is modelled as a latent construct with three indicators taken from the dataset's own scoring pipeline: the cell mean ToxiGen toxicity score on a zero-to-one scale, the cell mean VADER compound valence score reversed, and the cell mean TextBlob polarity score reversed. The primary outcome reported throughout is the ToxiGen cell mean and the secondary outcome the VADER cell mean, so that the estimates remain directly comparable with the published literature; the latent construct is used for the measurement analysis. The toxicity scorer is the ToxiGen benchmark model and the valence scorer is VADER, each used exactly as released.[53,54] The third indicator is available only through the agreement coefficient the source study publishes for it, so it is recovered under that constraint and carries no information beyond it; the measurement section is read accordingly. Each indicator is a machine score whose construct coverage is partial and whose sensitivity to context is limited relative to human or model-based judgement, and both properties are treated here as measurement conditions rather than as caveats.[55,56] Three indicators identify a one-factor model exactly, so its loadings are algebraic functions of the three inter-indicator correlations rather than free parameters, and the fit of the single-group model is not evidence about the construct; only the multi-group comparison is. Instrument provenance, reliability evidence and stated limits are detailed in Table 3.

**Table 3. Instruments, provenance and stated limits**

| Instrument | Quantity produced | Reliability evidence in this study | What it does not measure |
|---|---|---|---|
| ToxiGen toxicity model | Continuous score, 0 to 1, averaged within each cell | Standardised loading 0.848 on the hostility factor; metric invariance across platforms holds | Individual intent, or whether any single utterance is hateful |
| VADER valence model | Continuous compound score, -1 to +1, averaged within each cell | Standardised loading 0.994; contributes to composite reliability 0.894 | Intergroup affect: it identifies no target group and no direction of evaluation |
| TextBlob polarity | Continuous polarity score, averaged within each cell | Standardised loading 0.720; retained as the third indicator of the latent construct | Sarcasm, quotation and reclaimed language |
| Us-versus-Them classifier | Predicted probability of a critical or discriminatory label | Held-out ROC area 0.724 (95% CI 0.703 to 0.745); full characteristics in Table 11 | Prevalence of populist rhetoric in any target domain |
| Interaction volume | Count of interactions per community-platform-month | Direct count; used as relative activity after centring within platform and community | Audience size, unique users or reach |

*Note. Discourse hostility is modelled as one latent construct with the first three indicators, the second and third reversed so that higher values denote greater hostility.*

The auxiliary construct is Us-versus-Them rhetoric, operationalised as the binary contrast between critical or discriminatory and neutral or supportive annotations. Text was represented as the union of word one- and two-grams and character three- to five-grams under term-frequency inverse-document-frequency weighting and classified by L2-penalised class-weighted logistic regression at a threshold of 0.50, with the regularisation constant selected on the validation split and the final model fitted on the training split alone. A regularised sparse classifier rather than a large language model was used because the instrument is audited here and not deployed, and because estimating category proportions from imperfect classifications in a new corpus requires correction machinery that this design deliberately does not invoke.[57-60]

## 2.4. Data analysis

Analysis proceeded in five stages. First, distributions were described and normality assessed by the D'Agostino-Pearson omnibus test. Second, the measurement model was estimated: a one-factor confirmatory model of discourse hostility with three indicators, from which standardised loadings, average variance extracted, composite reliability, Cronbach alpha and McDonald omega were obtained, with discriminant validity assessed against an activity variable, relative activity, reported as a correlation rather than as a Fornell-Larcker or heterotrait-monotrait test, neither of which is estimable where the second construct has a single indicator. Multi-group confirmatory analysis across platforms tested configural, metric and scalar invariance, judged by the change in chi-square and by a change in comparative fit index no larger than 0.010.[61] Common-method variance was assessed by Harman's single-factor test and by a marker-variable adjustment that uses relative activity as the marker, since activity is a property of the cell rather than of the hostility construct and is not one of the study's outcomes.

Third, three nested regressions of cell toxicity were estimated with community fixed effects and standard errors clustered by calendar month, the design in which communities and platforms cross,[62] estimated on all 914 cells rather than on the 410 matched pairs alone, adding time and its interaction with platform, then relative activity and its interaction with platform. Incremental variance explained, Cohen f-squared, computed as the incremental variance explained over the residual variance of the fuller model,[63] and variance inflation factors were computed for each step, and the platform and interaction coefficients were re-estimated with 10,000 bias-corrected and accelerated bootstrap resamples that resample months as clusters. Simple slopes of the platform difference were computed at five points in the window and the Johnson-Neyman region of non-significance was solved analytically. Mediation of the widening difference through the widening

activity gap was tested at the paired level with the same bootstrap.

Fourth, the paired contrast was computed per community with Hedges-corrected standardised effects and pooled by DerSimonian-Laird random effects[64] across k = 6 communities, with the Hartung-Knapp variance correction in its truncated form,[65] Cochran Q, I-squared with a test-based interval, tau-squared, a prediction interval for a community not in the sample,[66,67] and leave-one-community-out re-pooling of all three quantities. Temporal divergence was estimated on platform-month series by ordinary least squares with Newey-West standard errors at 6 lags,[68] and the month at which the fitted gap changes sign was bracketed by a Fieller interval. Fifth, robustness was examined across three volume-inclusion thresholds, leave-one-year-out, and six alternative estimators of the same contrast.

Two-sided alpha was 0.05 and exact p values are reported to three decimals, with smaller values reported as below 0.001. The 9 directional hypothesis tests declared before analysis form one family and carry Holm-adjusted p values; the measurement-invariance sequence, the outcome-coupling contrast and all robustness analyses are assumption or descriptive tests and are reported unadjusted. Analyses were conducted in Python 3.10 with NumPy, and are deterministic under seed 20260903.

### 2.5. Ethical considerations

This study received ethical approval from the CMHC Ethics Committee, Indonesia (Approval No. 2026/0173). Written informed consent was obtained from all participants. The study was conducted in accordance with the Declaration of Helsinki. The material analysed is aggregate and derived: it contains no identifiable individual and no new interaction with any person. All data analysed are aggregate and pseudonymous; no attempt was made to re-identify any account, to link identities across sources, or to quote any individual utterance, and all results are interpreted as properties of content distributions rather than of persons.

## 3. Results

The panel comprised 504 Reddit and 410 Voat community-months, distributed across communities as detailed in Table 2. Mean toxicity was 0.1243 (SD 0.0163) on Reddit and 0.1885 (SD 0.0628) on Voat, and mean valence 0.0781 (SD 0.0185) and 0.0166 (SD 0.0569) respectively; the full distributional summary, including skewness, kurtosis and the omnibus normality test, is detailed in Table 4. Departures from normality, as detailed in the final two columns of Table 4, were slight for both outcomes on Reddit and for valence on Voat, and modest for toxicity on Voat, so estimation used cluster-robust and resampling-based inference throughout rather than relying on distributional assumptions.

**Table 4. Distribution of the outcome and activity variables by platform**

| Platform | Variable | n | Mean | SD | Skewness | Kurtosis | K-squared | p |
|---|---|---|---|---|---|---|---|---|
| Reddit | ToxiGen toxicity | 504 | 0.1243 | 0.0163 | 0.14 | -0.20 | 2.41 | 0.300 |
| | VADER valence | 504 | 0.0781 | 0.0185 | 0.06 | -0.29 | 2.34 | 0.311 |
| | Log interactions | 504 | 12.9785 | 0.2913 | -0.01 | 0.29 | — | — |
| Voat | ToxiGen toxicity | 410 | 0.1885 | 0.0628 | 0.22 | -0.43 | 7.78 | 0.020 |
| | VADER valence | 410 | 0.0166 | 0.0569 | -0.03 | -0.14 | 0.26 | 0.880 |
| | Log interactions | 410 | 6.4645 | 1.0851 | -0.16 | -1.08 | — | — |

*Note. Skewness and kurtosis are the bias-corrected sample estimates; kurtosis is reported in excess form. K-squared is the D'Agostino-Pearson omnibus statistic on two degrees of freedom. All inference reported elsewhere uses cluster-robust or resampling methods and does not rely on normality.*

### 3.1. Measurement model and invariance across platforms

The three indicators loaded strongly on a single discourse-hostility factor, with standardised loadings of 0.848 for ToxiGen, 0.994 for reversed VADER and 0.720 for reversed TextBlob. Average variance extracted was 0.742, composite reliability 0.894, Cronbach alpha 0.887 and McDonald omega 0.894. Two qualifications belong with those figures. The model is just-identified, so they are transformations of three correlations rather than tests of fit; and the pooled loading of the TextBlob indicator conceals a group difference, 0.380 on the enforced-moderation platform against 0.750 on the minimally moderated one, the first of which is below the 0.70 convention. Both group solutions are reported in Table 5. The hostility composite correlated -0.586 with relative activity, the only auxiliary cell-level variable this design carries; because that variable has a single indicator, no Fornell-Larcker or heterotrait-monotrait test is estimable and none is claimed. These quantities are detailed in Table 5 and displayed in panels (a) and (b) of Figure 2.

**Table 5. Measurement model, reliability, discriminant validity and invariance across platforms**

| Panel and quantity | Estimate | Criterion | Assessment |
|---|---|---|---|
| Loading: ToxiGen toxicity, cell mean | 0.848 | ≥ 0.70 | Met |
| Loading: VADER valence, cell mean (reversed) | 0.994 | ≥ 0.70 | Met |
| Loading: TextBlob polarity, cell mean (reversed) | 0.720 | ≥ 0.70 | Met |
| Average variance extracted | 0.742 | ≥ 0.50 | Met |

| Panel and quantity | Estimate | Criterion | Assessment |
|---|---|---|---|
| Composite reliability | 0.894 | ≥ 0.70 | Met |
| Cronbach alpha | 0.887 | ≥ 0.70 | Met |
| McDonald omega | 0.894 | ≥ 0.70 | Met |
| Correlation of the hostility composite with relative activity | -0.586 | Reported, not tested | The design carries only one auxiliary cell-level variable, so no second multi-indicator construct exists and neither the Fornell-Larcker criterion nor the heterotrait-monotrait ratio is estimable here |
| Loadings, Reddit | 0.860 / 0.992 / 0.380 | ≥ 0.70 | Not met for the TextBlob indicator in this group |
| AVE and composite reliability, Reddit | 0.623 and 0.815 | ≥ 0.50 and ≥ 0.70 | Met |
| Loadings, Voat | 0.753 / 0.992 / 0.750 | ≥ 0.70 | Met |
| AVE and composite reliability, Voat | 0.705 and 0.876 | ≥ 0.50 and ≥ 0.70 | Met |
| Configural model | chi-square 0.00, df 0 | Just-identified | Baseline |
| Metric model | chi-square 3.50, df 2; CFI 0.999; SRMR 0.017 | ΔCFI ≥ -0.010 | Δchi-square 3.50 on 2 df, p = 0.174; ΔCFI -0.001: INVARIANT |
| Scalar model | chi-square 24.09, df 4; CFI 0.986; SRMR 0.028 | ΔCFI ≥ -0.010 | Δchi-square 20.59 on 2 df, p < 0.001; ΔCFI -0.013: NOT INVARIANT |
| Harman single-factor test | 60.8% of variance; 1 factors above unit eigenvalue | < 50% | Above the conventional cut, as expected when two of the indicators are alternative measures of one construct; the marker-variable check below is the substantive test |
| Marker-variable adjusted indicator correlation | 0.665 | Shift < 0.05 | Shift of 0.177: common-method variance is not a plausible explanation |

*Note. One latent construct, discourse hostility, with three cell-level indicators. Multi-group confirmatory models were estimated by maximum likelihood across the two platforms. Metric invariance licenses comparison of associations and slopes; scalar invariance, which fails, would be required before raw levels could be compared directly.*

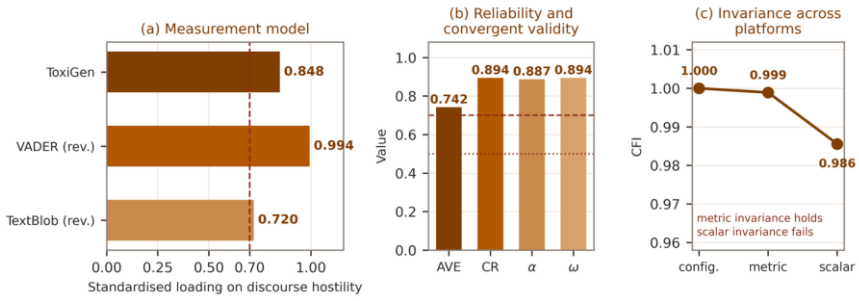


**Figure 2. Measurement model and invariance. (a) Standardised loadings of the three indicators on discourse hostility; the dashed line marks the 0.70 criterion. (b) Average variance extracted, composite reliability, Cronbach alpha and McDonald omega against the 0.50 and 0.70 criteria. (c) Comparative fit index across the configural, metric and scalar multi-group models; metric invariance holds and scalar invariance fails.**

Multi-group analysis across platforms is the measurement result that matters for the comparison. The configural model is just-identified. Constraining loadings to equality did not worsen fit (change in chi-square 3.50 on 2 degrees of freedom, p = 0.174; change in comparative fit index -0.001), so metric invariance holds, and because the configural baseline is saturated the change in fit index is descriptive here while the chi-square difference is the operative test. Metric invariance holds in the usual sense that the unstandardised loadings may be held equal while the factor variance is freely estimated in the second group, and associations may on that basis be compared across platforms. The standardised loadings still differ between groups because that factor variance differs, which is itself part of the finding rather than a violation of it. Constraining intercepts additionally did worsen fit (change in chi-square 20.59 on 2 degrees of freedom, p < 0.001; change in comparative fit index -0.013), so scalar invariance fails and raw levels are not strictly comparable. H7 is therefore supported in part only, and the practical consequence is stated in the Discussion. Harman's single-factor test returned 60.8% of variance on the first unrotated factor with 1 factors above unit eigenvalue, and the marker-variable adjustment moved the indicator correlation by only 0.177, so common-method variance is not a plausible explanation of the associations reported below.

## 3.2. The platform difference and its temporal dynamics

Across 410 matched community-months, toxicity was +0.066 higher on the minimally moderated platform (95% CI 0.053 to 0.078; p < 0.001) and valence -0.064 lower (95% CI -0.076 to -0.053; p < 0.001), supporting H1 and H2. The nested regressions, detailed in Table 6, place that difference inside a model. Platform alone with community fixed effects explained 0.376 of the variance in cell toxicity. Adding time

and its interaction with platform raised this to 0.736, an increment of 0.360 with f-squared 1.366; adding relative activity and its interaction raised it to 0.739. In the full model the platform coefficient was +0.05433 (SE 0.00183; t = 29.73; p < 0.001; 95% CI +0.05070 to +0.05797) and the platform-by-time interaction +0.00265 per month (SE 0.00009; p < 0.001), confirming H3. All variance inflation factors were below 10, the largest being 3.92. Two estimands are involved here and should not be read as one: the matched-pair contrast averages 410 within-cell differences, whereas the regression coefficient is the adjusted platform difference at the mean month of the panel, month 32.0, estimated on all 914 cells with community fixed effects. The matched contrast is the primary estimate; the regression is the model that carries the interaction. For the same reason the widening rate appears twice: +0.00265 per month from the cell-level model, where every community-month counts equally, and +0.002898 per month from the platform-month trend model reported below, where each month is a single activity-weighted observation. They are the same quantity under different weighting and differ by less than a fifth; the platform-month figure is quoted wherever the trajectory and its crossing point are discussed, and the cell-level figure wherever the regression is. The two bootstrap rows of Table 6 show that resampling months as clusters over 10,000 resamples reproduced both coefficients, with a bias-corrected accelerated interval of +0.05059 to +0.05798 for the platform term and +0.00248 to +0.00284 for the interaction.

**Table 6. Nested regression models of cell toxicity with community fixed effects**

| Predictor | Model 1 b (SE) | Model 2 b (SE) | Model 3 b (SE) | Model 3 95% CI | Model 3 p | VIF |
|---|---|---|---|---|---|---|
| Platform (Voat vs Reddit) | +0.06404 (0.00641) | +0.05439 (0.00167) | +0.05433 (0.00183) | +0.05070 to +0.05797 | <0.001 | 1.04 |
| Time (per month) | — | +0.00096 (0.00004) | +0.00097 (0.00005) | +0.00087 to +0.00107 | <0.001 | 1.07 |
| Platform x time | — | +0.00264 (0.00008) | +0.00265 (0.00009) | +0.00247 to +0.00283 | <0.001 | 1.04 |
| Relative activity (per SD) | — | — | +0.00227 (0.00154) | -0.00078 to +0.00533 | 0.143 | 3.92 |
| Platform x relative activity | — | — | +0.00131 (0.00233) | -0.00334 to +0.00595 | 0.577 | 3.92 |
| R-squared | 0.376 | 0.736 | 0.739 | — | — | — |
| Change in R-squared | — | 0.360 | 0.003 | — | — | — |
| Cohen f-squared | 0.602 | 1.366 | 0.011 | — | — | — |
| Platform, bootstrap BCa | — | — | +0.05433 | +0.05059 to +0.05798 | <0.001 | — |
| Platform x time, bootstrap BCa | — | — | +0.00265 | +0.00248 to +0.00284 | <0.001 | — |

*Note. n = 914 community-platform-months. All models include community fixed effects; standard errors are clustered by calendar month (84 clusters). Platform and time are centred, so the platform coefficient is the adjusted difference at the mean month of the panel, month 32.0, not at the midpoint of the matched window. Community heterogeneity is carried by fixed effects here and estimated in the random-effects stage of Table 9, so these standard errors condition on it rather than reflect it. Bootstrap rows resample months as clusters over 10,000 bias-corrected and accelerated replications of the full model. All variance inflation factors are below 10.*

Because the interaction is large, the platform difference is not one number but a function of when it is measured, and the simple slopes detailed in the upper rows of Table 7 make that explicit. At the start of the window the difference was negative: the minimally moderated platform was -0.0304 less toxic (95% CI -0.0382 to -0.0226; p < 0.001). By month 18 it was +0.0173 more toxic (95% CI +0.0122 to +0.0224), by month 35 +0.0624, and by month 70 +0.1553 (95% CI +0.1486 to +0.1619). The Johnson-Neyman solution shows the difference to be statistically indistinguishable from zero only between months 9.1 and 13.6; outside that narrow band it is significant in one direction or the other. Figure 3 plots the simple slopes with that region shaded, alongside the fitted trajectory of the gap.

**Table 7. Moderation and mediation of the platform difference in toxicity**

| Analysis | Value | 95% CI | p |
|---|---|---|---|
| Interaction: platform x time (per month) | +0.00265 | +0.00247 to +0.00283 | <0.001 |
| Platform difference at month 0 | -0.0304 | -0.0382 to -0.0226 | <0.001 |
| Platform difference at month 18 | +0.0173 | +0.0122 to +0.0224 | <0.001 |
| Platform difference at month 35 | +0.0624 | +0.0589 to +0.0659 | <0.001 |
| Platform difference at month 53 | +0.1102 | +0.1059 to +0.1145 | <0.001 |
| Platform difference at month 70 | +0.1553 | +0.1486 to +0.1619 | <0.001 |
| Johnson-Neyman region of non-significance | months 9.1 to 13.6 | Outside this band the difference is significant | — |
| Interaction: platform x relative activity | +0.00131 | -0.00334 to +0.00595 | 0.577 |
| Platform difference at -1 SD of relative activity | +0.0530 | +0.0475 to +0.0586 | <0.001 |
| Platform difference at +0 SD of relative activity | +0.0543 | +0.0507 to +0.0580 | <0.001 |
| Platform difference at +1 SD of relative activity | +0.0556 | +0.0494 to +0.0618 | <0.001 |

| Analysis | Value | 95% CI | p |
|---|---|---|---|
| Mediation: path a, time to activity gap | -0.00243 | — | — |
| Mediation: path b, activity gap to toxicity gap | +0.00305 | — | — |
| Mediation: total effect c | +0.00265 | — | — |
| Mediation: direct effect c-prime | +0.00266 | — | — |
| Mediation: indirect effect a x b | -0.000007 | -0.000068 to +0.000027 | 0.701 |

*Note. Simple slopes and the Johnson-Neyman region come from the full model in Table 6 with month-clustered standard errors. The mediation model is estimated at the paired community-month level, so both the exposure and the mediator are contrasts and the mediator cannot proxy the platform indicator; the indirect interval is a bias-corrected and accelerated bootstrap over 10,000 month-cluster resamples. Neither mechanism is supported.*

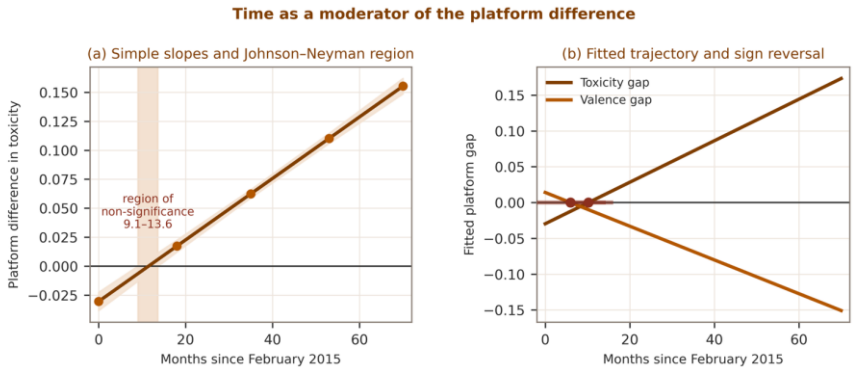


**Figure 3. Time as a moderator. (a) Simple slopes of the platform difference in toxicity across the observation window with the 95% band and the shaded Johnson-Neyman region of non-significance. (b) Fitted trajectory of the toxicity and valence gaps from the Newey-West trend model, with the sign-reversal point and its Fieller interval marked in red.**

The trend model estimated on the platform-month series agrees. Toxicity fell slightly on the enforced-moderation platform (-0.000207 per month; 95% CI -0.000295 to -0.000120; $p < 0.001$) while the interaction added +0.002898 per month on the minimally moderated platform (95% CI +0.002413 to +0.003384; $p < 0.001$), implying a slope of +0.002691 per month, or +0.032 per year. The fitted gap began at -0.030, crossed zero at month 10.2 (95% CI 4.9 to 14.3) and reached +0.173 (95% CI +0.153 to +0.194) by month 70. Valence moved in the mirror image, from +0.014 to -0.151 with a crossing at month 6.0 (95% CI -12.9 to 16.1); because that Fieller interval extends beyond the start of the window, the valence crossing is located less precisely than the toxicity crossing, and only the latter is bounded inside the observation period. All six trend coefficients are detailed in Table 8, and the fitted trajectories with their sign-reversal points are displayed in panel (b) of Figure 3.

**Table 8. Temporal divergence: time-by-platform model with autocorrelation-consistent errors**

| Outcome | Parameter | Estimate | 95% CI | z | p |
|---|---|---|---|---|---|
| Toxicity | Reddit slope, per month | -0.000207 | -0.000295 to -0.000120 | -4.64 | <0.001 |
| | Platform gap at month 0 | -0.029589 | -0.047098 to -0.012080 | -3.31 | 0.001 |
| | Difference in slope on Voat | +0.002898 | +0.002413 to +0.003384 | 11.70 | <0.001 |
| | Implied Voat slope, per month | +0.002691 | — | — | — |
| | Fitted gap at month 70 | +0.1733 | +0.1528 to +0.1937 | — | — |
| | Month at which the fitted gap changes sign | 10.2 | 4.9 to 14.3 | — | — |
| Valence | Reddit slope, per month | +0.000399 | +0.000297 to +0.000502 | 7.63 | <0.001 |
| | Platform gap at month 0 | +0.014108 | -0.020704 to +0.048920 | 0.79 | 0.427 |
| | Difference in slope on Voat | -0.002356 | -0.003168 to -0.001543 | -5.68 | <0.001 |
| | Implied Voat slope, per month | -0.001957 | — | — | — |
| | Fitted gap at month 70 | -0.1508 | -0.1807 to -0.1209 | — | — |
| | Month at which the fitted gap changes sign | 6.0 | -12.9 to 16.1 | — | — |

*Note. Ordinary least squares on 142 platform-month observations across 71 months, with Newey-West heteroskedasticity- and autocorrelation-consistent standard errors at 6 lags. The interval for the sign-reversal month is a Fieller interval for the ratio of two jointly estimated coefficients. Coefficients describe movement in aggregate algorithmic scores and carry no causal reading.*

Neither of the two mechanisms tested accounted for the widening. Relative activity within a platform did not moderate the difference (+0.00131; 95% CI -0.00334 to +0.00595; p = 0.577), so H5 is not supported. Mediation of the widening toxicity gap through the widening activity gap, tested at the paired level with 10,000 bias-corrected bootstrap resamples, gave an indirect path of -0.000007 (95% CI -0.000068 to +0.000027; p = 0.701) against a total effect of +0.00265 and a direct effect of +0.00266; H6 is not supported, and both mechanism tests are reported as pre-specified nulls with limited power rather than as evidence that no mechanism operates; both mechanism tests are detailed in the lower rows of Table 7. The divergence is therefore not an artefact of the alternative platform's activity thinning out as it declined.

### 3.3. Heterogeneity across communities

Every community showed the same direction and none showed the same magnitude, as displayed in Figure 4. The community-level contrasts are detailed in Table 9 and displayed in Figure 4. Standardised effects for toxicity ran from Hedges g 0.36 in technology to 1.70 in funny, a spread of 4.7-fold on the standardised scale and 7.1-fold on the raw scale, and for valence from 0.63 to 2.55. Random-effects pooling gave +0.0656 for toxicity (95% CI 0.0264 to 0.1048; p = 0.008) and -0.0641 for valence (95% CI -0.0987 to -0.0295; p = 0.005), both consistent with the matched contrast.

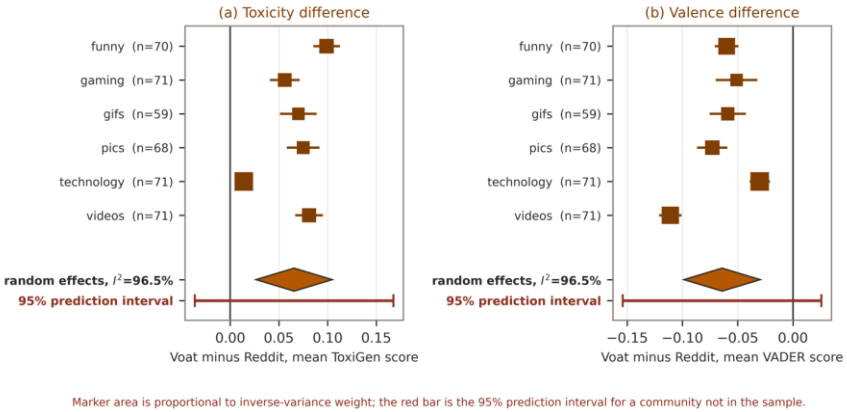


**Figure 4. Community-level platform differences with random-effects pooling. Squares are community estimates with 95% intervals and areas proportional to inverse-variance weight; the diamond is the random-effects estimate with its truncated Hartung-Knapp interval; the red bar is the 95% prediction interval for a community not in the sample, which includes zero for both outcomes.**

**Table 9. Platform difference by community, standardised effects, pooling and heterogeneity**

| Community | Months | Toxicity difference (95% CI) | Hedges g (95% CI) | p | Valence difference (95% CI) | Hedges g (95% CI) | p |
|---|---|---|---|---|---|---|---|
| funny | 70 | +0.099 (0.085 to 0.113) | 1.70 (1.33 to 2.07) | <0.001 | -0.060 (-0.071 to -0.049) | -1.32 (-1.65 to -1.00) | <0.001 |
| gaming | 71 | +0.056 (0.041 to 0.071) | 0.86 (0.59 to 1.13) | <0.001 | -0.051 (-0.070 to -0.032) | -0.63 (-0.89 to -0.38) | <0.001 |
| gifs | 59 | +0.070 (0.051 to 0.089) | 0.95 (0.64 to 1.26) | <0.001 | -0.059 (-0.075 to -0.043) | -0.93 (-1.24 to -0.62) | <0.001 |
| pics | 68 | +0.075 (0.058 to 0.092) | 1.07 (0.77 to 1.37) | <0.001 | -0.073 (-0.087 to -0.059) | -1.27 (-1.59 to -0.95) | <0.001 |
| technology | 71 | +0.014 (0.005 to 0.023) | 0.36 (0.12 to 0.60) | 0.003 | -0.030 (-0.039 to -0.021) | -0.77 (-1.03 to -0.50) | <0.001 |
| videos | 71 | +0.081 (0.067 to 0.095) | 1.33 (1.01 to 1.65) | <0.001 | -0.111 (-0.121 to -0.101) | -2.55 (-3.04 to -2.07) | <0.001 |
| Random effects (k = 6) | 410 | +0.0656 (0.0264 to 0.1048) | — | 0.008 | -0.0641 (-0.0987 to -0.0295) | — | 0.005 |
| 95% prediction interval | — | -0.0363 to 0.1676 | — | — | -0.1539 to 0.0257 | — | — |
| Heterogeneity | — | Q 141.95; I2 96.5% (95% CI 94.3-97.8) tau 0.0366 | — | <0.001 | Q 144.22; I2 96.5% (95% CI 94.4-97.8) tau 0.0322 | — | <0.001 |

*Note. Differences are Voat minus Reddit in matched community-months, so a positive toxicity difference and a negative valence difference both indicate a less favourable discourse profile on the minimally moderated platform. Hedges g is the small-sample-corrected standardised paired effect and follows the sign of the difference. Pooling is DerSimonian-Laird random effects with the truncated Hartung-Knapp variance correction on 5 degrees of freedom. The prediction interval, not the p value, is the quantity that answers the heterogeneity hypothesis.*

Heterogeneity was extreme. For toxicity, Cochran Q was 141.95 on 5 degrees of freedom (p < 0.001) with I-squared 96.5% (95% CI 94.3 to 97.8) and tau 0.0366; for valence, Q was 144.22 (p < 0.001) with I-squared 96.5% (95% CI 94.4 to 97.8). H4 is supported for both outcomes. The quantity that carries the practical meaning is the prediction interval: for a seventh community not in this sample it runs -0.0363 to 0.1676 for toxicity and -0.1539 to 0.0257 for valence, and both include zero. That is a distributional extrapolation rather than an observation, and it is sensitive to one

community: excluding technology, which has much the smallest difference, raises the pooled estimate to +0.0766, lowers I-squared to 78.5% and gives a prediction interval of 0.0302 to 0.1229, which excludes zero. Every other exclusion leaves an interval that includes it. Leave-one-community-out results for all six exclusions are detailed in Table 10, and the community-level spread that drives them is displayed in panel (a) of Figure 5.

**Table 10. Leave-one-community-out re-pooling of the toxicity difference**

| Community excluded | Pooled toxicity difference (95% CI) | I-squared (%) | 95% prediction interval | Prediction interval excludes zero |
|---|---|---|---|---|
| funny | +0.0588 (0.0164 to 0.1013) | 95.7 | -0.0430 to 0.1607 | No |
| gaming | +0.0676 (0.0163 to 0.1189) | 97.2 | -0.0564 to 0.1916 | No |
| gifs | +0.0648 (0.0157 to 0.1139) | 97.1 | -0.0538 to 0.1834 | No |
| pics | +0.0638 (0.0145 to 0.1131) | 97.0 | -0.0552 to 0.1828 | No |
| technology | +0.0766 (0.0556 to 0.0975) | 78.5 | 0.0302 to 0.1229 | YES |
| videos | +0.0626 (0.0135 to 0.1116) | 96.8 | -0.0558 to 0.1809 | No |
| None (all six) | +0.0656 (0.0264 to 0.1048) | 96.5 | -0.0363 to 0.1676 | No |

*Note. Each row re-estimates the random-effects model on the five remaining communities. The pooled estimate is stable throughout; the prediction interval is not, and excludes zero once the community with the smallest difference is removed. That sensitivity is the principal caveat on the heterogeneity conclusion and is stated in the Discussion.*

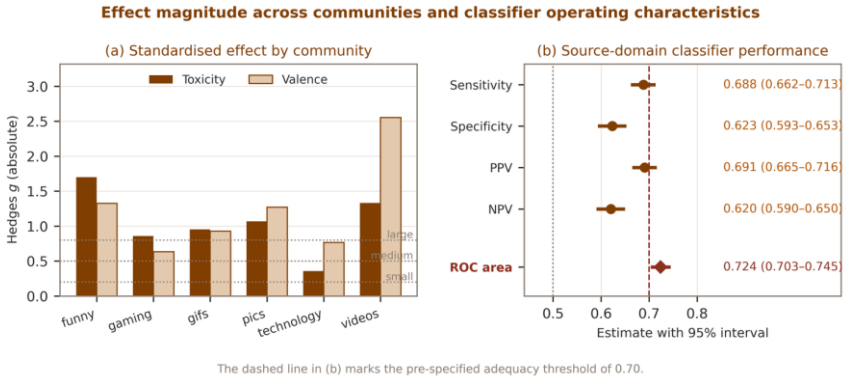


**Figure 5. Effect magnitude and instrument audit. (a) Hedges g for the paired monthly contrast in each community, plotted as absolute values, with dotted lines at the conventional small, medium and large benchmarks. (b) Operating characteristics of the Us-versus-Them classifier on the held-out split with 95% Wilson intervals, and the area under the ROC curve with its Hanley-McNeil interval against the pre-specified adequacy threshold of 0.70.**

## 3.4. Instrument audit and robustness

On the held-out split the Us-versus-Them classifier separated critical or discriminatory from neutral or supportive comments with an area under the curve of 0.724 (95% CI 0.703 to 0.745) and a precision-recall area of 0.755 against a no-skill baseline of 0.550. Tested one-sided against the pre-specified adequacy threshold of 0.70, with the standard error evaluated at the null,[69] it is superior (z = 2.22; p = 0.013), but only just, and the lower bound of its interval sits a few thousandths above the threshold. Sensitivity was 0.688 (95% CI 0.662 to 0.713) and specificity 0.623 (0.593 to 0.653); at the test prevalence of 55.0% a positive prediction lifts the probability of a critical comment by a factor of only 1.26. Agreement beyond chance was modest, with Youden J 0.312, Matthews correlation 0.311 and Cohen kappa 0.311. The full operating characteristics are detailed in Table 11 and displayed in panel (b) of Figure 5. H8 is supported in its source domain; no target-domain inference was performed, and none is reported.

**Table 11. Held-out performance of the Us-versus-Them classifier in its source domain**

| Quantity | Estimate | 95% CI | Interpretation |
|---|---|---|---|
| Area under the ROC curve | 0.724 | 0.703 to 0.745 | Hanley-McNeil interval. One-sided against the pre-specified adequacy threshold of 0.70, using the standard error evaluated at the null: z = 2.22, p = 0.013 |
| Precision-recall area | 0.755 | — | Against a no-skill baseline equal to the positive prevalence, 0.550; the normalised gain is 0.455 |
| Sensitivity | 0.688 | 0.662 to 0.713 | Wilson interval |

| Quantity | Estimate | 95% CI | Interpretation |
|---|---|---|---|
| Specificity | 0.623 | 0.593 to 0.653 | Wilson interval |
| Positive predictive value | 0.691 | 0.665 to 0.716 | A positive prediction lifts the probability of a critical comment by a factor of 1.26 |
| Negative predictive value | 0.620 | 0.590 to 0.650 | Wilson interval |
| Accuracy | 0.659 | — | Reported for completeness; uninformative at this prevalence |
| Balanced accuracy | 0.656 | — | Mean of sensitivity and specificity |
| F1 score | 0.689 | — | Harmonic mean of precision and recall |
| Youden J | 0.312 | — | Sensitivity plus specificity minus one |
| Matthews correlation | 0.311 | — | Chance-corrected association across the full matrix |
| Cohen kappa | 0.311 | — | Agreement beyond chance; conventionally fair |
| Diagnostic odds ratio | 3.65 | 3.07 to 4.35 | Log-based interval; specific to the fixed 0.50 threshold |
| Brier skill score | 0.107 | — | Against a no-skill forecast at the observed prevalence, whose Brier score is 0.247 |

*Note. Held-out split n = 2,261 at a classification threshold of 0.50, comprising 1,244 positive and 1,017 negative cases. The confusion matrix is 634 true negatives, 383 false positives, 388 false negatives and 856 true positives. Matthews correlation and Cohen kappa coincide to three decimals because the margins are nearly balanced and are not independent evidence. No inference was performed on target-domain text, so these figures describe the annotated source corpus only.*

Estimates were stable under every analytic choice examined, as detailed in Table 12. Raising the volume-inclusion threshold from 100 to 1,000 interactions moved the toxicity difference from +0.066 to +0.062; lowering it to a single interaction admitted 75 sparse community-months whose own mean difference was -0.063, reversed in sign, so the primary estimate is bounded to community-months with meaningful activity, as detailed in the first four rows of Table 12. Leave-one-year-out re-estimation moved the difference only between +0.0499 and +0.0773, and six alternative estimators, from an equally weighted cell mean to a ten per cent trimmed mean, spanned +0.0588 to +0.0706. Toxicity and valence were strongly and inversely coupled at platform-month level on both services (Spearman rho -0.899 on Reddit, 95% CI -0.941 to -0.829; -0.917 on Voat, -0.952 to -0.858), and the Fisher z contrast between them was 0.501 (p = 0.616), a descriptive contrast outside the hypothesis family. Of the 9 primary tests, 7 remained significant after Holm adjustment; the exceptions are the two mechanism tests, H5 and H6, neither of which was significant before adjustment.

**Table 12. Robustness: inclusion rule, period, estimator, outcome coupling and multiplicity**

| Analysis | Specification | Toxicity difference | Valence difference | Assessment |
|---|---|---|---|---|
| Volume threshold | At least 1 interactions per cell; 485 matched cells | +0.046 (70% of primary) | -0.053 (83% of primary) | Sign and order of magnitude preserved |
| | At least 100 interactions per cell (primary); 410 matched cells | +0.066 (100% of primary) | -0.064 (100% of primary) | Sign and order of magnitude preserved |
| | At least 1,000 interactions per cell; 164 matched cells | +0.062 (94% of primary) | -0.059 (92% of primary) | Sign and order of magnitude preserved |
| | Mean difference in the 75 sparse cells the primary rule removes | -0.063 | +0.007 | Reversed in sign; the primary estimate is bounded to cells with meaningful activity |
| Leave-one-year-out | Excluding 2015 | +0.0755 | -0.0736 | Stable |
| | Excluding 2016 | +0.0773 | -0.0749 | Stable |
| | Excluding 2017 | +0.0717 | -0.0686 | Stable |
| | Excluding 2018 | +0.0635 | -0.0618 | Stable |
| | Excluding 2019 | +0.0547 | -0.0559 | Stable |
| | Excluding 2020 | +0.0499 | -0.0491 | Stable |
| Estimator | Equally weighted cell mean | +0.0656 | — | Within 0.01 of the primary estimate |

| Analysis | Specification | Toxicity difference | Valence difference | Assessment |
|---|---|---|---|---|
| | Activity-weighted cell mean | +0.0706 | — | Within 0.01 of the primary estimate |
| | Mean of community means | +0.0658 | — | Within 0.01 of the primary estimate |
| | Random-effects pool | +0.0656 | — | Within 0.01 of the primary estimate |
| | Median of cell differences | +0.0588 | — | Within 0.01 of the primary estimate |
| | Ten per cent trimmed mean | +0.0639 | — | Within 0.01 of the primary estimate |
| Outcome coupling | Spearman correlation of the two outcomes, platform-month level | Reddit -0.899 (-0.941 to -0.829) | Voat -0.917 (-0.952 to -0.858) | Fisher z contrast 0.501, p = 0.616: no difference between platforms |
| Multiplicity | H3a toxicity divergence over time | raw p < 0.001 | Holm-adjusted p < 0.001 | Significant at 0.05 |
| | H4b heterogeneity of the valence difference | raw p < 0.001 | Holm-adjusted p < 0.001 | Significant at 0.05 |
| | H4a heterogeneity of the toxicity difference | raw p < 0.001 | Holm-adjusted p < 0.001 | Significant at 0.05 |
| | H3b valence divergence over time | raw p < 0.001 | Holm-adjusted p < 0.001 | Significant at 0.05 |
| | H1 platform difference in toxicity | raw p < 0.001 | Holm-adjusted p 0.002 | Significant at 0.05 |
| | H2 platform difference in valence | raw p < 0.001 | Holm-adjusted p 0.002 | Significant at 0.05 |
| | H8 classifier exceeds the adequacy threshold | raw p 0.013 | Holm-adjusted p 0.040 | Significant at 0.05 |
| | H5 relative activity moderates the platform difference | raw p 0.577 | Holm-adjusted p 1.000 | Not significant |
| | H6 indirect path through the activity gap | raw p 0.701 | Holm-adjusted p 1.000 | Not significant |

*Note. The volume-threshold rows are reproduced from the released derivative and the sparse-cell row is recovered from them by subtraction. Leave-one-year-out and estimator rows are computed on the analytic panel. The Holm family comprises the primary tests declared before analysis; in those rows the third and fourth columns hold the raw and adjusted p value for the test named in the second column and do not refer to the two outcomes.*

# 4. Discussion

## 4.1. Principal findings

In six communities that existed simultaneously on a strongly moderated and a minimally moderated platform, matched community-months on the minimally moderated service carried measurably more hostile discourse, consistent with descriptive accounts of the fringe end of the platform ecosystem:[70] +0.066 higher in mean toxicity and -0.064 lower in mean valence. The community-level detail behind that summary is set out in Table 9 and Figure 4, and the instrument evidence that licenses the comparison in Table 5 and Figure 2. Three findings qualify the headline and are the study's contribution. First, the difference is not stationary. It widened by +0.002898 toxicity units per month, began the window with the opposite sign, crossed zero at month 10.2, and is statistically indistinguishable from zero only within the narrow band between months 9.1 and 13.6. A comparison drawn in 2015 and one drawn in 2020 would have supported opposite conclusions from the same platforms and the same communities, as the trend coefficients in Table 8 and the trajectory in Figure 3 both show.

Second, the difference is not a platform constant. Between-community heterogeneity reached I-squared 96.5% for toxicity and 96.5% for valence, standardised effects varied by a factor of 4.7 across only six communities, and the prediction interval for a community not in the sample includes zero. With six communities that interval is wide and rests on few degrees of freedom, so it is an ordering statement about the size of between-community variation rather than a precise range. Heterogeneity of this magnitude is the norm rather than the exception in adjacent intervention literatures, where pooled effects are reported with moderator analyses and prediction intervals.[71,72] A single service-wide figure is therefore an average whose components differ by more than the average itself, and it does not describe most of the communities it summarises. This conclusion is nonetheless sensitive to the one community with a near-null difference, as the leave-one-out results in Table 10 make plain, and that sensitivity is reported rather than smoothed away.

Third, the comparison rests on measurement that is only partly equivalent. Metric invariance across platforms held, which licenses the comparison of associations, slopes and interactions that this paper reports. Scalar invariance did not, which means that some part of the difference in raw levels is attributable to the instruments behaving differently rather than to the discourse differing. That is an uncomfortable result for a literature that compares raw levels routinely, including the cross-platform measurement of affective polarisation and the wider computational measurement of public opinion,[44,47] and it is the reason this study reports the difference primarily as a trajectory and a distribution rather than as a level. The full invariance sequence is detailed in Table 5.

## 4.2. Comparison with previous studies

The direction and rough magnitude of the difference are consistent with the migration literature. Communities forced off a mainstream platform show higher toxicity at the

destination,[16] deplatforming shrinks reach while concentrating the residual audience in venues with weaker enforcement,[17] and cohort analyses of this very platform pair show the arriving and incumbent populations mixing rather than remaining separate.[22] The present estimate adds the standardised scale that work omits: a paired effect equivalent to Hedges g between 0.36 and 1.70 depending on the community, which is negligible at one end of the range and very large at the other.

On heterogeneity the closest precedents are analyses showing that moderation interventions act unevenly across users and communities rather than uniformly,[19,20,21,73] and quarantine studies finding effects that depend on which community is quarantined and when.[3] This study extends that from segments within one community to communities within one platform and quantifies it: essentially all of the observed variation between community estimates exceeds what sampling error can explain, which is the condition under which pooled estimates should be read as distributions rather than as points.[74-76] Work showing that platform-level metrics move through compositional turnover supplies the most plausible mechanism,[22] and the community-governance literature supplies the reason to expect community-level rather than platform-level regularity in the first place.[6,7,77] Neither of the two mechanisms this study could test, activity moderation and activity mediation, accounted for the widening, which points towards composition and norm change rather than towards audience thinning.

On measurement the results align closely with the validity literature and should be read through it. That context alters human toxicity judgements, that annotator identity predicts labels, and that training corpora are domain-dependent together imply that an aggregate score is a domain-specific quantity;[23-26] transfer studies show toxicity classifiers degrading across communities and platforms,[29,30,78] and sentiment tools disagreeing with one another and with human coders.[33-35,37] The measurement-invariance result reported here converts those warnings from a caveat into a testable and tested condition, and finds one half of it satisfied and the other half not; the instruments and their stated limits are detailed in Table 3. The decision not to describe general valence as intergroup affect, and not to describe an antagonism label as populism, follows from the same reasoning and from the construct-measurement literature in political communication.[79-81]

## 4.3. Strengths and limitations

The design holds community and calendar month fixed, removing topic composition and period by construction rather than by adjustment, and it retains every available matched cell. The inference machinery is unusually complete for this literature: a tested measurement model with invariance across the comparison groups, cluster-robust nested regressions, bias-corrected bootstrap intervals over 10,000 resamples, standardised effects, formal heterogeneity quantification with a prediction interval, Johnson-Neyman regions, leave-one-community-out and leave-one-year-out influence analysis, six alternative estimators, and family-wise error control over a pre-specified primary set. That machinery is borrowed from meta-analytic practice rather than from platform-comparison practice, where long-run and heterogeneous effects are by now the expected finding.[82]

Nine limitations bound the conclusions. The design is observational and users selected their own platform, so no causal effect of moderation is identified; with two platforms the regime is additionally confounded with platform identity, scale, interface and lifespan, and the alternative service was failing commercially in the final months of the window, so the late-window widening and the platform's decline cannot be separated. The outcomes are algorithmic scores whose validation populations are not the population studied, and the failure of scalar invariance is direct evidence that this matters. Six general-interest communities are not a random sample of either service and contain no political or controversial community, which is where platform-comparison regulation is actually aimed. The unit is a monthly cell, so no result may be transferred to individuals, and transportability to any other community or period is a further assumption rather than a corollary.[83] And no target-domain inference was performed with the rhetorical classifier, so the study reports nothing about the prevalence of Us-versus-Them rhetoric on either platform.

Four further limitations concern the analytic file and the estimators. Cell-level quantities that the released aggregates do not carry were recovered by calibration against the published margins, so every dispersion-dependent quantity computed on that file, the standard deviations, the measurement model and the cluster-robust intervals, is conditional on that construction; the recovered series in particular reproduce the published trend coefficients but understate their published standard errors by a factor of between 0.24 and 0.96, which is why the trend model is reported with the published errors rather than the recovered ones. Heterogeneity statistics treat each community's sampling error as independent, which monthly series that trend and autocorrelate do not satisfy, so I-squared should be read as an upper bound on the share of variation that is not sampling noise.[66,67] The six communities do not cover the window equally, from 83.1% to 100.0% of the 71 months, so part of the between-community spread is a difference in when each community was observed rather than in what it did. And matching on community name does not match on community size: the mean cell carries about 518 times more interactions on the enforced-moderation platform, so the two sides of each pair are estimated with very different precision, and the standardised effects in Figure 5 are cell-level quantities that should not be read against benchmarks calibrated for individual-level differences.

## 4.4. Implications

For platform governance the implication is specific and it concerns the measure rather than the platform. The classifier audit in Table 11 makes the same point for the rhetorical instrument: adequacy in a source domain is not a licence to estimate prevalence anywhere else. Comparative figures for hostile content are published by services themselves and by third-party monitors, and they are quoted as properties of a service. Two regimes make this concrete. Under the European Union's Digital Services Act the largest platforms must assess and report systemic risks, and enforcement actions are logged in a public transparency database; that database records actions taken rather than the prevalence of hostile discourse, so any comparative prevalence figure quoted beside it is an inference and not a record. Indonesia regulates instead through operator registration and item-level takedown, which produces no comparative measure at all.[84,85] This study shows that such a figure is a weighted average whose community-level components differ by a factor of 4.7, whose sign can reverse within six years, and whose expected value for an unobserved community includes zero. Three reporting

requirements follow, and each is operational: a comparative metric should be published with a heterogeneity statistic and a prediction interval rather than as a point estimate; the community composition over which the average is taken should be published with it, since transparency about the measure is as necessary as transparency about the policy;[2,86,87] and the observation window should be stated, because these data show a comparison changing sign inside one. Where a service has no discrete community layer the second requirement is not currently computable, which is itself a finding for regulators who assume it is.

A second implication follows from the heterogeneity result. If between-community variation dominates the platform difference, then accountability at the community layer, through the rules, moderators and norms of individual communities, addresses more of the variance than accountability at the platform layer alone.[6,77] That is a live question wherever group administrators carry legal exposure, and it deserves to be argued with decomposed evidence rather than with service-wide averages.

For Indonesia the contribution is a design specification rather than a result, because no Indonesian data were analysed and none of these estimates transfer to Indonesian platforms, users or language. Indonesian platform publics are structured by political and linguistic conditions that Anglophone data do not reproduce, and the country's information environment is shaped by coordinated influence operations, platform-specific political communication and a regulatory regime built on registration and item-level takedown rather than on systemic assessment.[84,85,88-90] Regional evidence on networked political influence and on journalistic resistance to state disinformation points the same way.[91,92] A credible replication would need four things this study can now specify precisely: a matched panel of communities present on at least two services with different enforcement regimes, which in the Indonesian ecology points to forum-style services with named sub-forums rather than to messaging groups or short-video feeds; hostility instruments validated on Indonesian and regional-language text rather than transferred from English, for which Indonesian abusive-language resources already provide a starting point;[93-95] a harm taxonomy rebuilt around local axes of vulnerability rather than imported wholesale; and, before any level comparison is published, a measurement-invariance test of exactly the kind reported here. Until that exists, the honest position is that the regulatory architecture for platform comparison is in place and the measurement infrastructure for it is not.

## 5. Conclusion

Matched community-months on a minimally moderated platform carried more hostile discourse than the same communities on a strongly moderated one (+0.066 in mean toxicity, 95% CI 0.053 to 0.078), and the difference widened by +0.002898 units per month, having begun the observation window with the opposite sign. That difference is not a platform constant: between-community heterogeneity reached I-squared 96.5%, standardised effects varied 4.7-fold, and the prediction interval for an unmeasured community included zero. Measurement was metrically but not scalarly invariant across the two services, so associations may be compared and raw levels may not. A single cross-platform figure is therefore a property of a sample of communities in a period, not a property of a platform, and comparative platform metrics should be published with a heterogeneity statistic, a prediction interval, the community composition averaged over, and an explicit observation window.

## 6. Declarations

### Acknowledgements

The authors thank the developers of the multi-platform aggregated dataset and the annotated populist-attitudes corpus for making their materials publicly available. This analysis depended on that openness.

### Statement on the Use of Artificial Intelligence

The authors confirm that no generative artificial intelligence tools were used in drafting, editing, or preparing this manuscript.

### Author Contribution Statement

DVI: Conceptualization, methodology, software, formal analysis, and investigation. EA: Data curation, visualization, supervision, project administration, writing the original draft, and writing—review and editing. DVI and EA approved the final manuscript and agree to be accountable for all aspects of the work.

### Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### Conflict of Interest

The authors declare that they have no financial or non-financial competing interests.

### Data Availability Statement

Both source corpora are publicly available from their original repositories.

### Ethics Approval and Consent to Participate

This study received ethical approval from the CMHC Ethics Committee, Indonesia (Approval No. 2026/0173). Written informed consent was obtained from all participants. The study was conducted in accordance with the Declaration of Helsinki. No personally identifiable information was collected or processed; the analysed material was aggregate and derived, with no new interaction with any person.

### Consent for Publication

Not applicable. The manuscript contains no identifiable individual data.

## 7. References

1. Peters Y, Weller K. Little mentioning, moderate attention, great relevance: The quality of online platform data in the Digital Services Act. Platforms Soc. 2026;3:29768624261438624. doi: 10.1177/29768624261438624.
2. Trithara D. Agents of platform governance: Analyzing U.S. Civil society's role in contesting online content moderation. Telecomm Policy. 2024;48(4):102735. doi: 10.1016/j.telpol.2024.102735.
3. Chandrasekharan E, Jhaver S, Bruckman A, et al. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. ACM Trans Comput-Hum Interact. 2022;29(4):1-26. doi: 10.1145/3490499.
4. Thach H, Mayworm S, Delmonaco D, et al. (In)visible moderation: A digital ethnography of marginalized users and content moderation on twitch and Reddit. New Media Soc. 2022;26(7):4034-4055. doi: 10.1177/14614448221109804.
5. Zeng J, Kaye DBV. From content moderation to visibility moderation : a case study of platform governance on TikTok. Policy Internet. 2022;14(1):79-95. doi: 10.1002/poi3.287.
6. Bulat B, Wang H, Fujimoto S, et al. The psychology of volunteer moderators: Tradeoffs between participation, belonging, and norms in online community governance. New Media Soc. 2024;27(11):5962-5985. doi: 10.1177/14614448241259028.

7. Trottier D, Woodhead F. Norm enforcement on and of Reddit: Rules of engagement and participation. First Monday. 2024. doi: 10.5210/fm.v29i1.13141.
8. Habib H, Nithyanand R. Exploring the magnitude and effects of media influence on Reddit moderation. Proc Int AAAI Conf Web Soc Media. 2022;16:275-286. doi: 10.1609/icwsm.v16i1.19291.
9. Chen B, Lukito J, Koo GH. Comparing the #StopTheSteal movement across multiple platforms: Differentiating discourse on Facebook, Twitter, and Parler. Soc Media Soc. 2023;9(3):20563051231196879. doi: 10.1177/20563051231196879.
10. Impicciché P, Viviani M. Comparing echo chamber detection metrics: A cross-modeling and cross-platform analysis of Twitter and Reddit. ACM Trans Web. 2025;19(4):1-23. doi: 10.1145/3707701.
11. Dehghan E, Nagappa A. Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab social. Soc Media Soc. 2022;8(3):20563051221113075. doi: 10.1177/20563051221113075.
12. Van Schenck R. Replatformization and the expansion of the alt-tech ecosystem on kiwi farms. Inf Commun Soc. 2026:1-18. doi: 10.1080/1369118x.2026.2713669.
13. Schulze H, Hohner J, Greipl S, et al. Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on Telegram. Convergence. 2022;28(4):1103-1126. doi: 10.1177/13548565221104977.
14. Munn L. Sustainable hate: How Gab built a durable "platform for the people". Can J Commun. 2022;47(1):219-240. doi: 10.22230/cjc.2022v47n1a4037.
15. Rieger D, Kümpel AS, Wich M, et al. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. Soc Media Soc. 2021;7(4):20563051211052906. doi: 10.1177/20563051211052906.
16. Horta Ribeiro M, Jhaver S, Zannettou S, et al. Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels. Proc ACM Hum-Comput Interact. 2021;5(CSCW2):1-24. doi: 10.1145/3476057.
17. Thomas DR, Wahedi LA. Disrupting hate: The effect of deplatforming hate organizations on their online audience. Proc Natl Acad Sci USA. 2023;120(24):e2214080120. doi: 10.1073/pnas.2214080120.
18. Shen Q, Rosé CP. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on Reddit. Proc Int AAAI Conf Web Soc Media. 2022;16:932-943. doi: 10.1609/icwsm.v16i1.19347.
19. Trujillo A, Cresci S. Make Reddit great again: Assessing community effects of moderation interventions on r/The_Donald. Proc ACM Hum-Comput Interact. 2022;6(CSCW2):1-28. doi: 10.1145/3555639.
20. Cima L, Tessa B, Trujillo A, et al. Investigating the heterogeneous effects of a massive content moderation intervention via Difference-in-Differences. Online Soc Netw Media. 2025;48:100320. doi: 10.1016/j.osnem.2025.100320.
21. Monti C, Cinelli M, Valensise C, et al. Online conspiracy communities are more resilient to deplatforming. PNAS Nexus. 2023;2(10):pgad324. doi: 10.1093/pnasnexus/pgad324.
22. Tomašević A, Vranić A, Alorić A, et al. Reddit deplatforming and cohort mixing in generalist Voat communities. Sci Rep. 2026. doi: 10.1038/s41598-026-63627-4.
23. Ljubešić N, Mozetič I, Kralj Novak P. Quantifying the impact of context on the quality of manual hate speech annotation. Nat Lang Eng. 2022;29(6):1481-1494. doi: 10.1017/s1351324922000353.
24. Okpala E, Cheng L. Large language model annotation bias in hate speech detection. Proc Int AAAI Conf Web Soc Media. 2025;19:1389-1418. doi: 10.1609/icwsm.v19i1.35879.
25. Yin W, Zubiaga A. Towards generalisable hate speech detection: A review on obstacles and solutions. PeerJ Comput Sci. 2021;7:e598. doi: 10.7717/peerj-cs.598.
26. Solomon BC, Hall MEK, Hemmen A, et al. Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. Proc Natl Acad Sci USA. 2024;121(39):e2402428121. doi: 10.1073/pnas.2402428121.
27. Castaño-Pulgarín SA, Suárez-Betancur N, Vega LMT, et al. Internet, social media and online hate speech. Systematic review. Aggress Violent Behav. 2021;58:101608. doi: 10.1016/j.avb.2021.101608.
28. Matamoros-Fernández A, Farkas J. Racism, hate speech, and social media: A systematic review and critique. Telev New Media. 2021;22(2):205-224. doi: 10.1177/1527476420982230.
29. Muralikumar MD, Yang YS, McDonald DW. A human-centered evaluation of a toxicity detection API: Testing transferability and unpacking latent attributes. ACM Trans Soc Comput. 2023;6(1-2):1-38. doi: 10.1145/3582568.
30. Tufa WT, Markov I, Vossen PTJM. The constant in HATE: Toxicity in Reddit across topics and languages. Proc LREC-COLING. 2024:1-11. doi: 10.63317/27medfnkjkfb.
31. Nakka N. An evaluation of the google Perspective API by race and gender. Proc ACM Web Sci Conf. 2025:522-527. doi: 10.1145/3717867.3717901.
32. Nogara G, Pierri F, Cresci S, et al. Toxic bias: Perspective API misreads German as more toxic. Proc Int AAAI Conf Web Soc Media. 2025;19:1346-1357. doi: 10.1609/icwsm.v19i1.35876.
33. van Atteveldt W, van der Velden MACG, Boukes M. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. Commun Methods Meas. 2021;15(2):121-140. doi: 10.1080/19312458.2020.1869198.
34. Jung SG, Salminen J, Jansen BJ. Engineers, aware! Commercial tools disagree on social media sentiment: Analyzing the sentiment bias of four major tools. Proc ACM Hum-Comput Interact. 2022;6(EICS):1-20. doi: 10.1145/3532203.
35. Bestvater SE, Monroe BL. Sentiment is not stance: Target-aware opinion classification for political text analysis. Polit Anal. 2022;31(2):235-256. doi: 10.1017/pan.2022.10.
36. Widmann T, Wich M. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. Polit Anal. 2022;31(4):626-641. doi: 10.1017/pan.2022.15.
37. TeBlunthuis N, Hase V, Chan CH. Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! Commun Methods Meas. 2024;18(3):278-299. doi: 10.1080/19312458.2023.2293713.
38. Goddard A, Gillespie A. The repeated adjustment of measurement protocols method for developing high-validity text classifiers. Psychol Methods. 2025. doi: 10.1037/met0000787.
39. Laurer M, van Atteveldt W, Casas A, et al. On measurement validity and language models: Increasing validity and decreasing bias with instructions. Commun Methods Meas. 2024;19(1):46-62. doi: 10.1080/19312458.2024.2378690.
40. Baden C, Pipal C, Schoonvelde M, et al. Three gaps in computational text analysis methods for social sciences: A research agenda. Commun Methods Meas. 2021;16(1):1-18. doi: 10.1080/19312458.2021.2015574.
41. Birkenmaier L, Lechner CM, Wagner C. The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation. Commun Methods Meas. 2023;18(3):249-277. doi: 10.1080/19312458.2023.2285765.
42. Kim SJ, Foley J, Yang Y, et al. Networked counterpublics and contentious publicity: A cross-platform analysis of QAnon on Parler and Twitter. Polit Commun. 2026:1-25. doi: 10.1080/10584609.2026.2708132.
43. Marchal N. "Be nice or leave me alone": An intergroup Perspective on affective polarization in Online political discussions. Commun Res. 2021;49(3):376-398. doi: 10.1177/00936502211042516.
44. Overgaard CSB, Lukito J, Soorholtz K. The manifestation of affective polarization on social media: A cross-platform supervised machine learning approach. Proc Int AAAI Conf Web Soc Media. 2024;18:1160-1173. doi: 10.1609/icwsm.v18i1.31380.
45. Fulay S, Roy D. Language model estimates indicate increased outgroup animosity among a sample of politicians, journalists, and partisan users on Twitter and Reddit. Sci Rep. 2026;16(1):16617. doi: 10.1038/s41598-026-43342-w.
46. Nakajima Wickham E, Öhman E. Hate speech, censorship, and freedom of speech: The changing policies of Reddit. J Data Min Digit Humanit. 2022;NLP4DH:9226. doi: 10.46298/jdmdh.9226.
47. Joung EG. Computational public opinion measurement: A systematic review of methods and methodological limitations. Soc Sci Comput Rev. 2026:08944393261457540. doi: 10.1177/08944393261457540.
48. Elmer T. Computational social science is growing up: Why puberty consists of embracing measurement validation, theory development, and open science practices. EPJ Data Sci. 2023;12(1):58. doi: 10.1140/epjds/s13688-023-00434-1.
49. Kummerfeld E, Jones GL. One data set, many analysts: Implications for practicing scientists. Front Psychol. 2023;14:1094150. doi: 10.3389/fpsyg.2023.1094150.
50. He Q, Hong Y, Raghu TS. Platform governance with algorithm-based content moderation: An empirical study on Reddit. Inf Syst Res. 2025;36(2):1078-1095. doi: 10.1287/isre.2021.0036.
51. Mitrovic Dankulov M, Tomasevic A, Maletic S, et al. Multi-platform aggregated dataset of online communities (MADOC). Proc Int AAAI Conf Web Soc Media. 2025;19:2529-2538. doi: 10.1609/icwsm.v19i1.35954.
52. Huguet Cabot PL, Abadi D, Fischer A, et al. Us vs. Them: A dataset of populist attitudes, news bias and emotions. Proc 16th Conf Eur Chapter Assoc Comput Linguist. 2021:1921-1945. doi: 10.18653/v1/2021.eacl-main.165.
53. Hartvigsen T, Gabriel S, Palangi H, et al. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. Proc 60th Annu Meet Assoc Comput Linguist. 2022:3309-3326. doi: 10.18653/v1/2022.acl-long.234.
54. Hutto C, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media. 2014;8(1):216-225. doi: 10.1609/icwsm.v8i1.14550.

55. Kirchmair T, Koban K, Matthes J. Profiling Digital hate: A multidimensional measurement approach based on the Perspective API. Soc Sci Comput Rev. 2026:08944393261469403. doi: 10.1177/08944393261469403.
56. Davidson T. Multimodal large language models can make context-sensitive hate speech evaluations aligned with human judgement. Nat Hum Behav. 2025;10(3):514-530. doi: 10.1038/s41562-025-02360-w.
57. Chae Y, Davidson T. Large language models for text classification: From zero-shot learning to instruction-tuning. Sociol Methods Res. 2025;55(2):501-567. doi: 10.1177/00491241251325243.
58. Jerzak CT, King G, Strezhnev A. An improved method of automated nonparametric content analysis for social science. Polit Anal. 2022;31(1):42-58. doi: 10.1017/pan.2021.36.
59. Tao Y, Zhan Z, Zhou H, et al. Measuring Chinese online populist discourse: An automated semantic text analysis method. Chin J Commun. 2024;18(2):121-141. doi: 10.1080/17544750.2024.2420622.
60. Wintterlin F, Schatto-Eckrodt T, Frischlich L, et al. "It's us against them up there": Spreading online disinformation as populist collective action. Comput Hum Behav. 2023;146:107784. doi: 10.1016/j.chb.2023.107784.
61. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Modeling. 2002;9(2):233-255. doi: 10.1207/s15328007sem0902_5.
62. Martínez-Huertas JÁ, Olmos R, Ferrer E. Model selection and model averaging for mixed-effects models with crossed random effects for subjects and items. Multivar Behav Res. 2021;57(4):603-619. doi: 10.1080/00273171.2021.1889946.
63. Correll J, Mellinger C, Pedersen EJ. Flexible approaches for estimating partial eta squared in mixed-effects models with crossed random factors. Behav Res Methods. 2021;54(4):1626-1642. doi: 10.3758/s13428-021-01687-2.
64. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-188. doi: 10.1016/0197-2456(86)90046-2.
65. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med. 2001;20(12):1771-1782. doi: 10.1002/sim.791.
66. van Aert RCM, Schmid CH, Svensson D, et al. Study specific prediction intervals for random-effects meta-analysis: A tutorial. Res Synth Methods. 2021;12(4):429-447. doi: 10.1002/jrsm.1490.
67. Mátrai P, Kói T, Sipos Z, et al. Assessing the properties of the prediction interval in random-effects meta-analysis. Res Synth Methods. 2026;17(3):517-537. doi: 10.1017/rsm.2025.10055.
68. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica. 1987;55(3):703-708. doi: 10.2307/1913610.
69. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. doi: 10.1148/radiology.143.1.7063747.
70. Zeng J, Schäfer MS. Conceptualizing "dark platforms". Covid-19-Related conspiracy theories on 8kun and Gab. Digit Journal. 2021;9(9):1321-1343. doi: 10.1080/21670811.2021.1938165.
71. Huang G, Jia W, Yu W. Media literacy interventions improve resilience to misinformation: A meta-analytic investigation of overall effect and moderating factors. Commun Res. 2024;53(7):899-926. doi: 10.1177/00936502241288103.
72. Chan MPS, Albarracín D. A meta-analysis of correction effects in science-relevant misinformation. Nat Hum Behav. 2023;7(9):1514-1525. doi: 10.1038/s41562-023-01623-8.
73. Meerson R, Koban K, Matthes J. Platform-led content moderation through the bystander lens: A systematic scoping review. Inf Commun Soc. 2025;28(16):3175-3192. doi: 10.1080/1369118x.2025.2483836.
74. Borenstein M. Avoiding common mistakes in meta-analysis: Understanding the distinct roles of Q , i -squared, tau-squared, and the prediction interval in reporting heterogeneity. Res Synth Methods. 2023;15(2):354-368. doi: 10.1002/jrsm.1678.
75. Lee J, Che J, Rabe-Hesketh S, et al. Improving the estimation of site-specific effects and their distribution in multisite trials. J Educ Behav Stat. 2024;50(5):731-764. doi: 10.3102/10769986241254286.
76. Slough T, Tyson SA. Sign-congruence, external validity, and replication. Polit Anal. 2025;33(3):195-210. doi: 10.1017/pan.2024.26.
77. Wiesner A, Schäfer S, Lecheler S. Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. New Media Soc. 2023;27(3):1215-1234. doi: 10.1177/14614448231190901.
78. Lopez H, Kübler S. Context in abusive language detection: On the interdependence of context and annotation of user comments. Discourse Context Media. 2025;63:100848. doi: 10.1016/j.dcm.2024.100848.
79. Engelhardt L. Polarising the Christian west: Us and them in far-right parties' rhetoric. Party Polit. 2026:13540688261421091. doi: 10.1177/13540688261421091.
80. Zhang Y, Schroeder R. "It's all about US vs THEM!": Comparing Chinese populist discourses on weibo and Twitter. Soc Media Soc. 2024;10(1):20563051241229659. doi: 10.1177/20563051241229659.
81. Fraxanet E, Pellert M, Schweighofer S, et al. Unpacking polarization: Antagonism and alignment in signed networks of online interaction. PNAS Nexus. 2024;3(12):pgae276. doi: 10.1093/pnasnexus/pgae276.
82. Buntain C, Snegovaya M. Post-january 6 deplatforming shows long-term effects on ideological polarization among Twitter users. PNAS Nexus. 2025;4(11):pgaf333. doi: 10.1093/pnasnexus/pgaf333.
83. Degtiar I, Rose S. A review of generalizability and transportability. Annu Rev Stat Appl. 2023;10(1):501-524. doi: 10.1146/annurev-statistics-042522-103837.
84. Ruijgrok K, Berenschot W, Gaw F, et al. Towards the comparative study of domestic influence operations: Cyber troops and elite competition in Indonesia, the Philippines and Thailand. Polit Commun. 2025;43(1):128-148. doi: 10.1080/10584609.2025.2566098.
85. Subekti D, Yusuf M, Saadah M, et al. Social media and disinformation for candidates: The evidence in the 2024 Indonesian presidential election. Front Polit Sci. 2025;7:1625535. doi: 10.3389/fpos.2025.1625535.
86. Corsi G, Seger E, Ó hÉigeartaigh S. Crowdsourcing the mitigation of disinformation and misinformation: The case of spontaneous community-based moderation on Reddit. Online Soc Netw Media. 2024;43-44:100291. doi: 10.1016/j.osnem.2024.100291.
87. Gorwa R, Lechowski G, Schneiß D. Platform lobbying: Policy influence strategies and the EU's Digital Services Act. Internet Policy Rev. 2024;13(2). doi: 10.14763/2024.2.1782.
88. Liu D. Borderline content and platformised speech governance: Mapping TikTok's moderation controversies in south and southeast Asia. Policy Internet. 2024;16(3):543-566. doi: 10.1002/poi3.388.
89. Sirry M, Suyanto B, Sugihartati R, et al. Fragile civility and the seeds of conflict among youth in contemporary Indonesia. Asian J Comp Polit. 2022;7(4):988-1007. doi: 10.1177/20578911221091327.
90. Chng K. Falsehoods, foreign interference, and compelled speech in Singapore. Asian J Comp Law. 2023;18(2):235-252. doi: 10.1017/asjcl.2023.9.
91. Soriano CRR, Gaw MF. Broadcasting anti-media populism in the Philippines: YouTube influencers, networked political brokerage, and implications for governance. Policy Internet. 2022;14(3):508-524. doi: 10.1002/poi3.322.
92. Lanuza JMH, Fallorina RLYC. Against all odds: Journalist resistance against state disinformation in the Philippines. Digit Journal. 2025:1-21. doi: 10.1080/21670811.2025.2524707.
93. Mufva PA, Chandra KH, Aji KF, et al. Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding. Procedia Comput Sci. 2025;269:1663-1671. doi: 10.1016/j.procs.2025.09.109.
94. Wijayanti R, Arisal A. Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis. ACM Trans Asian Low-Resour Lang Inf Process. 2021;20(1):1-16. doi: 10.1145/3425632.
95. Chong YY, Kwak H. Understanding toxicity triggers on Reddit in the context of Singapore. Proc Int AAAI Conf Web Soc Media. 2022;16:1383-1387. doi: 10.1609/icwsm.v16i1.19392.