# Supplementary Information

Narrowing and recovery of aspect-sentiment diversity in online discussions

Xiaodan Wang[1,2], Yanbin Liu[3], Jianing Xia[4], Yining Wang[5], Minh Nguyen[2], Weihua Li[2*], Shiqing Wu[6*], Quan Bai[7]

[1]School of Engineering, Yanbian University, Yanji, Jilin Province, China.
[2]Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland, 1010, New Zealand.
[3]Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, 2007, Australia.
[3]Universal Higher Education, Melbourne, Australia.
[4]Faculty of Science, University of Melbourne, Melbourne, Australia.
[5]Faculty of Data Science, City University of Macau, Macau SAR, China.
[6]School of Information and Communication Technology, University of Tasmania, Hobart, 7001, Tasmania, Australia.

*Corresponding author(s). E-mail(s): weihua.li@aut.ac.nz; sqwu@cityu.edu.mo;

## Supplementary Methods

This supplement supports the main text with detailed definitions, uncertainty calculations, sensitivity analyses, provenance and the human audit of the model-generated labels. Supplementary Methods S1–S5 describe the primary $W = 30$ analyses, S6 describes uncertainty and alternative settings, S7 documents sample construction and cross-period comparability, and S8 describes the human label audit. Supplementary

Results report the lexical contrast, audit outcomes and schema-size sensitivity. Tables and figures follow their order of citation in the main text. Equation numbers refer to the main text. Null comparisons assess robustness to specified alternatives without redefining tunnel vision.

## S1. Thread, pair-space and window notation

Let thread $i$ contain $N_i$ usable comments in chronological order,

$$\mathcal{T}_i = (c_{i1}, \ldots, c_{iN_i}).$$

Its frozen event-specific aspect schema is

$$\mathcal{A}_i = \{a_{i1}, \ldots, a_{iK_i}\},$$

and the sentiment set is

$$\mathcal{S} = \{\text{positive}, \text{neutral}, \text{negative}\}.$$

The corresponding pair space is

$$\mathcal{P}_i = \mathcal{A}_i \times \mathcal{S},$$

with pair $p = (a, s)$ interpreted as one expressed evaluative position. This is a discussion-level representation of retained textual evaluations, not a direct measurement of participants' internal beliefs or cognitive states.

The fixed window length is $W = 30$ comments and the stride is $s = 5$. For window index $j = 1, \ldots, J_i$, let

$$e_{ij} = W + (j - 1)s$$

be the ending comment position and

$$\mathcal{W}_{ij} = \{c_{i,e_{ij}-W+1}, \ldots, c_{i,e_{ij}}\}.$$

If $n_{ij}(p)$ is the number of retained mentions of pair $p$ in the window, the total number of pair mentions is

$$M_{ij} = \sum_{p \in \mathcal{P}_i} n_{ij}(p),$$

and, for $M_{ij} > 0$, the empirical pair distribution is

$$q_{ij}(p) = \frac{n_{ij}(p)}{M_{ij}}.$$

A window with $M_{ij} = 0$ is missing rather than a zero-diversity observation. ASPE $H_{ij}$ is Eq. (1) of the main text and the equivalent diversity is $D_{ij} = 2^{H_{ij}}$.

For concentration, let $\mathcal{P}_{ij}^{(5)}$ contain the five most frequent pairs in $\mathcal{W}_{ij}$. Then

$$C_{ij}^{(5)} = \sum_{p \in \mathcal{P}_{ij}^{(5)}} q_{ij}(p).$$

The whole-thread trends are the Spearman correlations in Eq. (2) of the main text. A trend requires at least three valid windows. A constant series is assigned zero, and per-window values are rounded before ranking so that equal values tie deterministically.

## S2. Null randomisations and robustness descriptors

The density-preserving and order-shuffle nulls test different alternatives.

For the *density-preserving null*, let $m_{it}$ denote the number of retained pair mentions attached to comment position $t$ in thread $i$. The randomisation preserves the complete vector

$$(m_{i1}, \ldots, m_{iN_i})$$

and the whole-thread frequency $f_i(p)$ of every pair $p$. Conceptually, all pair labels are pooled and randomly dealt back into the fixed comment-position slots of sizes $m_{it}$. Hence the number of identifiable pairs carried by each comment position is unchanged while the identities of the pairs appearing at those positions are randomised.

For the *order-shuffle null*, the intact comment-level label bundle $B_{it}$, including an empty bundle for a comment with no retained pair, is preserved. A random permutation $\pi_b$ of the comment indices produces

$$(B_{i,\pi_b(1)}, \ldots, B_{i,\pi_b(N_i)}).$$

This preserves each comment's label bundle and the distribution of comment-level mention counts, but not the mention-count sequence at the original chronological positions.

Each null uses $B = 300$ permutations under fixed seeds. For null $\nu \in \{\text{den}, \text{ord}\}$, let $\rho^{\nu}_{ib}$ be the ASPE trend in permutation $b$. The one-sided Monte-Carlo probability is

$$p_i^{\nu} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(\rho^{\nu}_{ib} \le \rho_{\text{ASPE},i})}{B + 1}.$$

Let $O_i = 1$ if thread $i$ shows observed tunnel vision, that is $\rho_{\text{ASPE},i} < -0.15$ and $\rho_{\text{conc},i} > 0$, and $O_i = 0$ otherwise. Convenient indicator notation for the robustness descriptors is

$$Z_i^{\text{den}} = O_i\, \mathbf{1}(p_i^{\text{den}} < 0.05), \qquad Z_i^{\text{ord}} = O_i\, \mathbf{1}(p_i^{\text{ord}} < 0.05),$$

with

$$Z_i^{\mathrm{dual}} = Z_i^{\mathrm{den}} Z_i^{\mathrm{ord}}.$$

These descriptors overlap because they test different null hypotheses. Failure to meet one null criterion does not remove observed-tunnel-vision status. Robustness is conditional on each specified randomisation; these indicators are retrospective trajectory descriptors, not explanatory exposures for recovery.

## S3. Episode timing, competing endpoints and aligned trajectories

Let $r_{ij} = D_{ij}/D_{i,\mathrm{ref}}$ be the relative-diversity trajectory defined in Section 3.4 of the main text, where $D_{i,\mathrm{ref}}$ is the equivalent diversity of the first full window. Let $j_i^{\mathrm{on}}$ denote the first valid window index that begins a confirming run of three low windows:

$$j_i^{\mathrm{on}} = \min\left\{j : \; r_{ij}, r_{i,j+1}, r_{i,j+2} \leq 0.8\right\}.$$

The reported onset comment position is

$$\tau_i^{\mathrm{on}} = e_{i,j_i^{\mathrm{on}}},$$

and the relative onset position is

$$u_i^{\mathrm{on}} = \frac{\tau_i^{\mathrm{on}}}{N_i}.$$

Given an open episode, let

$$j_i^{\mathrm{rec}} = \min\left\{j > j_i^{\mathrm{on}} : \; r_{ij}, r_{i,j+1}, r_{i,j+2} \geq 0.9\right\}$$

be the first confirming recovery run, with

$$\tau_i^{\mathrm{rec}} = e_{i,j_i^{\mathrm{rec}}}.$$

For a recovered first episode,

$$T_i = \tau_i^{\mathrm{rec}} - \tau_i^{\mathrm{on}}.$$

If the observed record ends before recovery,

$$T_i = N_i - \tau_i^{\mathrm{on}}.$$

Onset and recovery are dated to the first window of the confirming run; confirmation occurs ten comments later because the stride is five comments. The onset/recovery band provides hysteresis, and a missing window resets the consecutive-window counter. Recovery restores the specified relative-diversity level, not necessarily the earlier pair set or distribution. Timing is measured in comments, and recovery, persistence and relapse refer to the available observed record. A subsequent onset after a confirmed recovery is a relapse.

For persistence, define post-onset follow-up

$$F_i = N_i - \tau_i^{\mathrm{on}}$$

and the low-window share

$$P_i^{\mathrm{low}} = \frac{\sum_{j \geq j_i^{\mathrm{on}}} \mathbf{1}(r_{ij} \leq 0.8)}{\#\{j \geq j_i^{\mathrm{on}} : r_{ij} \text{ valid}\}}.$$

The persistent-episode indicator is

$$I_i^{\mathrm{pers}} = \mathbf{1}\big(F_i \geq 30 \ \wedge \ P_i^{\mathrm{low}} \geq 0.70 \ \wedge \ r_{i,\mathrm{last}} \leq 0.8\big)\,,$$

where $r_{i,\mathrm{last}}$ is the relative diversity of the last full window. Threads with fewer than 30 comments of post-onset follow-up are treated as having insufficient follow-up for this classification.

For the first episode, define endpoint type

$$J_i = \begin{cases} 1, & \text{confirmed recovery occurs before the record ends,} \\ 0, & \text{the observed record ends without recovery.} \end{cases}$$

For group $g$, the empirical cumulative incidence functions shown in Figure 3 are

$$F_g^{\mathrm{rec}}(\ell) = \frac{1}{n_g} \sum_{i \in g} \mathbf{1}(T_i \leq \ell,\, J_i = 1),$$

and

$$F_g^{\mathrm{end}}(\ell) = \frac{1}{n_g} \sum_{i \in g} \mathbf{1}(T_i \leq \ell,\, J_i = 0).$$

The remaining threads at lag $\ell$ have experienced neither endpoint. This is the empirical competing-endpoint construction used in the analysis rather than a model-based cause-specific hazard estimator.

For onset-aligned trajectories, a window ending at comment $e_{ij}$ is assigned lag

$$\ell_{ij} = e_{ij} - \tau_i^{\mathrm{on}}.$$

At each lag, the plotted group trajectory is

$$m_g(\ell) = \text{median}_{i \in g}\{r_{ij} : \ell_{ij} = \ell\},$$

using only threads still observed at that lag. The interquartile band is calculated across contributing threads; pointwise 95% community-cluster bootstrap intervals describe uncertainty in the median. Summaries are drawn only where at least 30 threads contribute.

The thread-level minimum relative diversity among threads with an episode is

$$r_i^{\min} = \min_{j:\, r_{ij} \text{ valid}} r_{ij},$$

the minimum over all valid windows of the thread, not only those in the first episode, and the relative position of the minimum is the end position of the minimizing window divided by $N_i$. The end-state ratio is $r_{i,\text{last}}$ (last full window). The fraction of the observed diversity loss regained by the end is

$$G_i = \frac{r_{i,\text{last}} - r_i^{\min}}{1 - r_i^{\min}}.$$

## S4. Pair-level block dynamics

Each thread is partitioned into non-overlapping 30-comment blocks. Let $n_{ik}(p)$ be the number of mentions of pair $p$ in block $k$, and let

$$A_{ik} = \{p : n_{ik}(p) > 0\}$$

be its active pair set. Retention $\text{Ret}_{ik}$ is Eq. (4) of the main text.

The later block's active pairs are partitioned into three disjoint sets. Continued pairs are

$$A_{ik}^{\text{cont}} = A_{ik} \cap A_{i,k-1}.$$

Returning pairs are absent from the immediately preceding block but occurred earlier:

$$A_{ik}^{\text{return}} = A_{ik} \cap \left[ \left( \bigcup_{h<k-1} A_{ih} \right) \setminus A_{i,k-1} \right].$$

First-observed pairs have never appeared in an earlier block:

$$A_{ik}^{\text{first}} = A_{ik} \setminus \left( \bigcup_{h<k} A_{ih} \right).$$

If $M_{ik} = \sum_p n_{ik}(p)$, their mention shares are

$$S_{ik}^{x} = \frac{\sum_{p \in A_{ik}^{x}} n_{ik}(p)}{M_{ik}}, \qquad x \in \{\text{cont}, \text{return}, \text{first}\}.$$

These shares sum to one.

To measure redistribution within the same pair set across adjacent blocks, define the common set

$$C_{ik} = A_{i,k-1} \cap A_{ik}.$$

For $h \in \{k-1, k\}$, renormalise the block distribution over only this common set,

$$\widetilde{q}_{ih}^{(ik)}(p) = \frac{n_{ih}(p)}{\sum_{q \in C_{ik}} n_{ih}(q)}, \qquad p \in C_{ik},$$

and compute

$$H_{ih}^{(ik)} = -\sum_{p \in C_{ik}} \widetilde{q}_{ih}^{(ik)}(p) \log_2 \widetilde{q}_{ih}^{(ik)}(p).$$

The common-set entropy change is

$$\Delta H_{ik}^{\text{common}} = H_{ik}^{(ik)} - H_{i,k-1}^{(ik)}.$$

Positive values indicate that the positions surviving the transition are used more evenly in the later block. Transitions with an empty common set have no common-set entropy estimate.

Let $p_{i,k-1}^{*}$ be the dominant pair in the earlier block. Ties are broken first by the pair's cumulative count in the thread and then lexicographically. Its share change into block $k$ is

$$\Delta d_{ik} = q_{ik}(p_{i,k-1}^{*}) - q_{i,k-1}(p_{i,k-1}^{*}),$$

and a separate indicator records whether the same pair remains dominant in block $k$.

Blocks are aligned separately on the block containing the first onset and the block containing the first recovery. Event-related paired changes use Eq. (5) of the main text. Baseline transition values are first averaged within thread over all valid transitions and then averaged across threads so that longer threads do not receive greater point weight.

Because first-observed pairs become less common with absolute thread position, the density-robust versus all-other contrast for the onset block is also directly standardised. The two groups are compared at the same absolute block positions and the position-specific contrasts are weighted by the density-robust group's observed block-position distribution.

## S5. Opening and lexical analyses

The opening comprises the first 50 usable comments. Exactly five 30-comment windows with stride five lie wholly inside this interval. If $\mathcal{J}_i^{\text{open}}$ denotes those windows, opening diversity is

$$H_i^{\text{open}} = \frac{1}{5} \sum_{j \in \mathcal{J}_i^{\text{open}}} H_{ij}.$$

Opening mention density is the number of retained pair mentions per opening comment; the active-pair count and dominant-pair share are also recorded.

Post-opening outcomes use only windows whose first comment is comment 51 or later, so no outcome window straddles the opening boundary. Let $\mathcal{J}_i^{\text{post}}$ denote these windows. The post-opening trend is

$$\rho_{H,\text{post},i} = \text{Spearman}(j, H_{ij}), \qquad j \in \mathcal{J}_i^{\text{post}},$$

the post-opening change is

$$\Delta H_{\text{post},i} = H_{i,\text{late}} - H_{i,\text{post},1},$$

and the late level is the entropy $H_{i,\text{late}}$ of the last full window.

For $Y_i \in \{\rho_{H,\text{post},i}, H_{i,\text{late}}\}$, the adjusted association is estimated as

$$Y_i = \beta_1 z(H_i^{\text{open}}) + \beta_2 z(\log N_i) + \beta_3 z(K_i) + \beta_4 z(M_i^{\text{open}}) + \gamma_{g(i)} + \varepsilon_i,$$

where $K_i$ is schema size, $M_i^{\text{open}}$ is opening mention density, $\gamma_{g(i)}$ is a community fixed effect, and $z(\cdot)$ denotes standardisation to one sample standard deviation. Communities containing a single thread are dropped under the within transformation, leaving 2,958 threads across 75 communities. Coefficient intervals use community-clustered CR1 standard errors and a $t$ reference distribution with $G-1$ degrees of freedom, where $G$ is the number of retained communities. Pooled specifications are reported for comparison. These are retrospective associations rather than prospective or causal models.

For the lexical contrast, text in each analysis window is HTML-decoded, URLs are removed, and the remaining text is lower-cased. Tokens are runs of letters and digits; no stop-word removal or stemming is applied. From each eligible window, 100 tokens

are sampled without replacement in each of 20 fixed-seed draws. If $V_{ij}^{(b)}$ is the number of distinct types in draw $b$, lexical diversity is

$$L_{ij} = \frac{1}{20} \sum_{b=1}^{20} V_{ij}^{(b)}.$$

The lexical trend is

$$\rho_{L,i} = \text{Spearman}(j, L_{ij}),$$

and the analysis reports the cross-thread Spearman association between $\rho_{L,i}$ and $\rho_{\text{ASPE},i}$ (Supplementary Results R1 and Figure S6). This is a contrast with one lexical measure, not a validation of semantic diversity.

## S6. Uncertainty, sensitivity and second-period replication

Uncertainty in descriptive aggregate estimates is summarised with a community-cluster bootstrap. For each of $B_{\text{boot}} = 2{,}000$ resamples, communities are sampled with replacement and all their threads are carried together. The 2019 analysis uses 84 communities. The same draws are reused across statistics so that group contrasts and paired changes are formed within each resample. Unstandardised point estimates assign equal weight to eligible threads, and reported 95% intervals are percentile intervals. Interquartile ranges are computed across individual threads. Regression intervals use clustered standard errors as specified in S5. These procedures describe uncertainty conditional on the sampling and measurement design; they do not establish representativeness beyond it.

For block-transition statistics, event-related differences are formed within thread before aggregation. Baseline transition statistics first average each quantity over valid transitions within thread and then average across threads. For the onset-aligned curves, cluster-bootstrap resampling is applied to the median trajectory at each fixed lag.

The main 30-comment measurement window is fixed and is not varied. Sensitivity checks concern only the descriptive episode rules and the reference choice. Onset/recovery thresholds are changed from 0.80/0.90 to 0.75/0.85 and 0.85/0.95 (T1 and T2). R0 and R1 begin detection after comment 50, using the first full-window reference or the equivalent diversity implied by mean opening entropy, respectively. No alternative was used to select the primary result. The influence of the five largest communities is examined by dropping each in turn and re-aggregating the frozen results.

The second-period comparison reports raw, thread-equal-weight estimates and estimates standardised over the 70 communities shared by the analysed samples. For standardisation, each community receives its share of the 2019 sample restricted to these communities, with equal weights among threads within a community. Conditional statistics retain the relevant episode or transition denominator, and each period is resampled at community level. These comparisons address measured community composition, not all possible differences in selection or labelling. The second-period sampling and pipeline are documented in S7.

## S7. Sample provenance and period comparability

***The full 2019 processing corpus and frozen census.***

The full corpus contains 6,128 processed Reddit threads from April 2019. The census covered every thread with a platform-reported comment count of 300–600 in 126 curated debate and discussion communities, without further filtering. Of its 4,161 threads, 3,821 were processed, 338 having no usable comments and two lacking a processing record. The additional sample comprised 2,580 threads sampled under fixed quotas from other comment-count bands, of which 2,307 were scored. This sample combines 2,500 threads randomly selected under the band quotas with 80 additional selections of very long threads. The full corpus yielded 4,219 threads with at least 150 usable

comments, constituting the *reliably measurable corpus*. The primary analyses use the census because the other bands have different sampling fractions.

Within the full corpus, 840 processed census threads have fewer than 150 usable comments and are excluded from temporal analysis. This floor provides sufficient windows for temporal measurement. Of the remaining 2,981, 14 have 602–3,239 usable comments and are retained as *supplementary long threads*, without contributing to inference. Platform counts are snapshots at archiving and can differ from usable-comment totals in the monthly dump. The resulting *analysis sample* contains 2,967 threads from 84 communities and 1,133,533 usable comments, with median thread length 369 (interquartile range 317–443; range 151–596). These threads are the denominator for 2019 whole-thread prevalence; episode and transition analyses use their eligible subsets.

***The final 2018 sample.***

April 2018 was selected before its results were computed. The 2018 sampling frame uses submission-dump `num_comments` in $[300, 600]$, at least 150 usable comments, and quotas proportional to the 2019 sample over the 84 communities represented in the analysis sample. The final manifest records 1,035 threads against a target of 1,200, with shortages left unfilled across communities. The final resampling seed is `20260920`, with stable CRC32-derived community keys and the final thread list saved. The final sample reuses 898 frozen schemas and adds 137 newly induced schemas. All reported gate comparisons use this same final sample and these same schemas; earlier sampling-frame estimates are superseded.

***Schema induction and pair annotation.***

A two-pass procedure uses `gpt-5-mini` with minimal reasoning effort and JSON-constrained output. An induction pass reads 80 evenly sampled comments and proposes 8–15 evaluable aspects, each with a name, stance-neutral definition and lexical cues.

A critic pass reads a fresh sample of 80 comments and can add, split, merge or refine aspects before freezing the schema. Sampling spans the thread, making the representation retrospective. Duplicate aspects within a comment are collapsed, and every retained comment–aspect match receives a positive, neutral or negative label. In the analysis sample, 51% of comments contain at least one retained pair, the stream contains 1,158,184 pair mentions, and a thread uses a median of 30 distinct pairs. Before the census run, the classifier path was accepted after reproducing, on three anchor threads labelled comment by comment by `gpt-5-mini`, reference narrowing correlations to within 0.10 in $\rho_{\text{ASPE}}$ with the same sign and ordering. This process check is distinct from the independent human audit in S8.

***Shared components.***

The gate scores a comment–aspect combination if the lower-cased comment contains a lower-cased cue phrase or a whitespace-delimited aspect-name token longer than three characters, matched as substrings with punctuation retained. The gate implementation reproduces the census mention sets exactly on an audit of 300 threads (116,900 comment–aspect pairs). Relabelling 80 randomly selected census threads with the second-period classifier path produced 99.9% sentiment agreement on matched pairs. These are process-consistency checks, not estimates of annotation accuracy. The classifier checkpoint, retention of neutral labels, absence of a confidence cut-off, and measurement, null and episode code are shared. The model checkpoint is `yangheng/deberta-v3-base-absa-v1.1`; the classifier input combines the comment with the aspect name.

***Representation differences between periods.***

The pipeline reuses frozen relevance matching, the classifier checkpoint, analysis code and available schemas, while inducing schemas for new threads and constructing the second-period sampling frame. The shared schema-induction procedure is generative,

so repeated runs need not produce identical names, definitions or cues. These elements determine relevance matches, and sharing the gate does not ensure equivalent representations across threads or periods. The density-preserving null conditions on mention counts and pair totals; it does not correct differences in which pairs were recognised. The comparison therefore assesses recurrence of the main patterns under the documented sampling frame, with raw and community-standardised estimates, rather than a causal period effect or strict measurement equivalence.

***Communities.***

The census list was assembled from community names containing *politic*, *debate*, *feminis* or *climate* together with manually selected news, advice and issue-specific communities; the length-band sample used the same list. The 126 communities are: `ABoringDystopia`, `Abortiondebate`, `Albertapolitics`, `AmericanPolitics`, `AmItheAsshole`, `Anarchism`, `Android`, `antifeminists`, `antiwork`, `apple`, `AskHistorians`, `AskMen`, `askscience`, `AskTrumpSupporters`, `AskWomen`, `atheism`, `AustinPolitics`, `australia`, `AustralianPolitics`, `badeconomics`, `Bitcoin`, `canada`, `CanadaPolitics`, `changemyview`, `Christianity`, `ClimateActionPlan`, `ClimateOffensive`, `climateskeptics`, `collapse`, `Conservative`, `conspiracy`, `CryptoCurrency`, `dating_advice`, `DebateAChristian`, `DebateAnAtheist`, `DebateAVegan`, `DebateReligion`, `DebateVaccine`, `democrats`, `economicCollapse`, `EconomicInsights`, `Economics`, `energy`, `environment`, `europe`, `EverythingScience`, `exmuslim`, `Feminism`, `Feministpassdenied`, `FeMRADebates`, `Firearms`, `FontPolitics`, `Futurology`, `GCdebatesQT`, `geopolitics`, `guns`, `GunsAreCool`, `history`, `IAMALiberalFeminist`, `india`, `IndiaNonPolitical`, `internationalpolitics`, `investing`, `ireland`, `JoeRogan`, `JordanPeterson`, `KansasPolitics`, `LateStageCapitalism`, `legaladvice`, `Libertarian`, `MarylandPolitics`, `MensRights`, `mensrightsindia`, `MGTOW`, `neoliberal`, `NeutralPolitics`, `news`, `newzealand`, `nottheonion`, `Our_Politics`,

personalfinance, philosophy, PhilosophyMemes, PikabuPolitics, POLITIC, Political_Revolution, Political_Tumor, PoliticalCompassMemes, PoliticalDiscussion, PoliticalHumor, PoliticalHumour, PoliticalVideo, politicalwarfare, politics, privacy, prochoice, prolife, PurplePillDebate, relationship_advice, relationships, religion, Republican, rPolitics, science, Scotland, socialism, technology, teslamotors, Texas_Politics, TexasPolitics, The_Donald, TrueChristian, TrueReddit, TwoXChromosomes, u_BrainyFeminist, UKPoliticalComedy, ukpolitics, unitedkingdom, unpopularopinion, UpliftingNews, uspolitics, vegan, vegetarian, wallstreetbets, worldnews, worldpolitics.

## S8. Human audit of the model-generated labels

***Purpose.***

The aspect-sentiment labels used throughout the study are produced by the frozen pipeline (relevance gate and aspect-conditioned sentiment classifier; main text, Section 3.2). Apart from the acceptance test on three anchor threads, the labels were not human-checked before the analyses. The audit therefore asks two questions: how closely the pipeline's pair assignments agree with independent human judgement, and whether that agreement changes with position in the thread, since a pipeline that recognised fewer pairs in later comments could produce apparent narrowing.

***Sample.***

Threads were drawn with seeded generators within strata defined by period and by the whole-thread classification (density-robust tunnel vision or other threads); no local-episode, recovery or persistence outcome entered the selection. The first package holds 40 threads, 30 from the April 2019 analysis sample (15 with density-robust tunnel vision and 15 other threads) and 10 from April 2018 (5 and 5); the second holds 30 further April 2019 threads (15 and 15), one from each of 30 communities not represented in the

first. In every thread, 12 comments were drawn uniformly from the early, middle and late thirds of the usable comments (4 per third), from all usable comments rather than from comments with a retained pair. The two annotators worked through the thread sheets independently and in their own order, with the classification and comment position withheld. The audit set is the intersection of the threads each had completed when the audit closed: 29 of the 40 threads in the first package and 13 of the 30 in the second, giving 42 threads (18 with density-robust tunnel vision and 24 other threads, 8 of them from April 2018). In 39 threads both annotators completed all 12 comments; in three threads one comment was left incomplete by one annotator, so the audit set holds 501 rather than 504 comments: 405 from April 2019 (167 in density-robust threads and 238 in other threads) and 96 from April 2018 (48 and 48), with 168, 166 and 167 comments in the early, middle and late thirds; 250 carry no pipeline pair. Selection probabilities, seeds and candidate lists are archived with the packages.

***Annotation protocol.***

Two annotators labelled every comment independently under written guidelines. For each comment they received the thread title, the thread's frozen schema with the full definition of every aspect, the target comment and the three usable comments immediately preceding it in chronological order. The period, the thread's classification, the pipeline's labels, the relevance-gate rule and the comment's position in the thread were withheld, and comments within a thread were presented in random order. For every aspect of the schema the annotator recorded ABSENT, POSITIVE, NEUTRAL or NEGATIVE; a comment may address several aspects with different sentiments. Judgements were to be made by meaning, so that a keyword is neither necessary nor sufficient. Difficult cases were flagged (mixed evaluation, sarcasm, unclear attribution, insufficient context, evaluation outside the schema) and every flagged comment required a short evidence note.

***Reference labels.***

Inter-annotator agreement provides a reference benchmark for interpreting pipeline–human agreement; it is not an upper bound on the pipeline's performance. The pipeline is compared with each annotator separately and with the *consensus cells*, the comment–aspect cells to which both annotators assigned the same label (5,468 of 6,071 cells, 90.1%, of which 216 carry a sentiment). For evaluation of the pipeline against human consensus, the pipeline metrics are restricted to these cells and the 603 cells on which the annotators disagree are excluded rather than treated as errors; inter-annotator statistics are computed on all cells. Comments flagged as uninterpretable without further context are retained; in 99% of their cells both annotators and the pipeline record ABSENT.

***Scoring.***

Agreement between two label sources is summarised at three levels: the per-comment F1 between their sets of (aspect, sentiment) pairs, averaged over comments, with two empty sets counted as agreement; the share of comments with identical pair sets; and, at cell level, agreement on whether an aspect is addressed (Cohen's $\kappa$) and, where both sources assign a sentiment, agreement on its polarity. Within the consensus cells, three pipeline quantities are distinguished. *Detection recall* is the share of consensus pairs for which the pipeline records a pair of any polarity, and *detection precision* the share of the pipeline's pairs in those cells that are consensus pairs. *Polarity agreement* is the share of detected consensus pairs to which the pipeline assigns the annotators' polarity. *Exact-pair recall* and *exact-pair precision* require both aspect and polarity to match; exact-pair recall equals detection recall multiplied by polarity agreement. The quantity relevant to narrowing is detection by thread third: a recall that fell with position would indicate that part of the measured decline in diversity reflects declining recognition rather than declining expression. Intervals are 95% percentile intervals from

a thread-stratified bootstrap of 2,000 resamples. Results are reported in Supplementary Results R2 and Table S6.

# Supplementary Results

## R1. Evaluative and lexical trends are largely unrelated

The trend in rarefied lexical diversity (Supplementary Methods S5) is largely unrelated to the ASPE trend across the 2,967 analysis-sample threads: the cross-thread Spearman correlation between $\rho_{L,i}$ and $\rho_{\mathrm{ASPE},i}$ is $+0.002$ (95% community-cluster bootstrap interval $[-0.037, 0.040]$; Figure S6). The median lexical trend among threads with density-robust tunnel vision is close to zero. Evaluative narrowing therefore need not coincide with declining vocabulary diversity under this measure. The contrast supports treating lexical richness and the distribution of aspect-sentiment positions as distinct empirical quantities; it does not validate the semantic representation by itself.

## R2. Human audit of the model-generated labels

Agreement is reported at three levels: *detection*, whether a comment addresses an aspect at all; *polarity*, the sentiment assigned when both sources record a pair; and *position*, whether detection changes across the early, middle and late thirds of a thread; exact-pair agreement, which requires both aspect and polarity to match, combines the first two. Table S6 follows the same order. Agreement between the two annotators provides the reference benchmark. On the 501 audited comments their pair sets have a mean F1 of 0.39 and are identical for 28% of comments; they agree on whether an aspect is addressed in 91% of cells ($\kappa = 0.48$) and, where both assign a sentiment, on its polarity in 74% of cells. Disagreement is therefore concentrated on whether a comment addresses an aspect at all, not on the direction of the evaluation. It is largest for comments flagged as sarcastic, where the annotators' mean F1 is 0.17 against 0.40 elsewhere: deciding what a sarcastic comment evaluates, and whose evaluation it conveys, divides human readers as much as it defeats the classifier. Disagreement of this kind is a documented property of evaluative annotation rather than a defect of these annotators: intended sarcasm is frequently missed by third-party readers [1], label

variation on subjective judgements persists among trained annotators [2], and in direct comparisons both humans and large language models show their lowest agreement on sarcasm [3].

*Detection.* Pipeline–human agreement falls within the range of this benchmark (Table S6). Its pair sets have a mean F1 of 0.65 with one annotator and 0.32 with the other, and they are identical to at least one annotator's for 60% of comments; this criterion allows a match with either annotator and is therefore not comparable with the 28% of comments on which the two annotators' sets coincide. At cell level it agrees with the two annotators on whether an aspect is addressed in 94% and 91% of cells ($\kappa = 0.62$ and 0.41), against 91% ($\kappa = 0.48$) between the annotators. Within the consensus cells it detects 169 of the 216 consensus pairs at aspect level (detection recall 0.78, 95% interval 0.70–0.85); 169 of its 301 detections in those cells are consensus pairs (detection precision 0.56, 0.45–0.68), and over all cells 65% of its pairs match at least one annotator. As for the annotators, its errors lie in whether an aspect is addressed rather than in polarity.

*Polarity.* Where the pipeline detects a consensus pair (169 pairs), it assigns the annotators' polarity in 96% of cases (162 pairs). This figure is conditional on cells where the annotators agree and is therefore not comparable with the 74% polarity agreement between the annotators, which is computed on the 290 cells where both record a pair of any polarity. Combining detection and polarity, 162 of the 216 consensus pairs are reproduced exactly (exact-pair recall 0.75, 0.66–0.82; exact-pair precision 0.54, 0.43–0.65). Of the 216 consensus pairs the pipeline misses 47 at aspect level and assigns a different polarity to a further 7; the misses are more often positive pairs (67% of positive consensus pairs are recovered with the correct label, against 77% of negative ones), and it assigns negative polarity more often than the annotators (63% of its pairs against 54–55% of theirs).

*Position.* The audit finds no evidence of declining detection with position in the thread. Detection recall is 0.74 (0.59–0.86) in the early third, 0.79 (0.67–0.88) in the middle third and 0.81 (0.69–0.91) in the late third; the early-minus-late difference is −0.08 with an interval (−0.22 to +0.06) that includes zero, and of the 47 consensus pairs the pipeline misses, 18 fall in the early third, 13 in the middle third and 16 in the late third. Exact-pair recall is 0.68 (0.54–0.80), 0.78 (0.64–0.88) and 0.79 (0.64–0.90) in the three thirds, with an early-minus-late difference of −0.11 (−0.26 to +0.05). Detection precision shows the same pattern: 0.55 (0.42–0.69), 0.51 (0.37–0.67) and 0.62 (0.49–0.74) in the three thirds, with an early-minus-late difference of −0.07 (−0.18 to +0.05); the pipeline's pairs in consensus cells that neither annotator recorded number 41, 48 and 43 by third. These estimates provide no evidence of a systematic late decline in recognition or a late rise in spurious pairs, but do not exclude measurement error in the estimated diversity trajectories. Detection recall is higher in density-robust threads (0.83) than in other threads (0.75), while detection precision is lower (0.46 against 0.67) and has similar point estimates in the early and late thirds of density-robust threads (0.46 and 0.47): these estimates do not show a substantial early-to-late change in precision within that group. Although the density-preserving null conditions on retained mention counts at each position, it cannot correct errors in pair identity or sentiment. In the April 2018 threads detection recall and precision are both 0.93 and exact-pair recall is 0.91, so the pipeline does not label the second period less faithfully than the first.

*Thread level.* Two further observations support the measurement at the level at which it is used. The number of distinct pairs recorded over a thread's twelve audited comments is more strongly correlated between the pipeline and each annotator (Spearman 0.55 and 0.76) than between the annotators (0.28). And the share of negative pairs rises from the early to the late third for both annotators (47% to 63% and 54% to 65%) as it does for the pipeline (52% to 75%): the growing weight of negative positions

in later comments, visible in the case threads of main-text Figure 5, is seen by human readers as well, although the pipeline overstates it.

## R3. Schema size and diversity measurement

The frozen schema size $K_i$ has a median of 12 aspects (interquartile range 12–13; range 3–15). The bands 3–10, 11, 12, 13 and 14–15 contain 77, 421, 1,304, 1,015 and 150 threads, respectively. We assess associations with the whole-thread measures using Spearman correlations and 95% intervals from 2,000 community-cluster bootstrap resamples.

Schema size is weakly associated with the ASPE trend ($\rho = 0.035$, interval $[-0.019, 0.075]$) and retained equivalent diversity ($\rho = 0.034$, $[-0.008, 0.075]$). The density-robust shares across the five bands are 7.8%, 10.0%, 10.4%, 7.6% and 10.0%, with overlapping intervals. Median schema size is 12 in both density-robust threads and all other threads.

Entropy levels show stronger associations with schema size. Early-window ASPE has a correlation of 0.199 ($[0.153, 0.238]$), and late-window ASPE a correlation of 0.220 ($[0.171, 0.262]$). After division by the pair-space capacity $\log_2(3K_i)$, they fall to 0.060 ($[0.010, 0.103]$) and 0.094 ($[0.046, 0.138]$), respectively. The opening models include schema size as a covariate (S5).

Capacity normalisation leaves within-thread ASPE ranks unchanged, including in the null permutations. It therefore preserves ASPE rank trends and the associated whole-thread classifications, with concentration measured as before. It does not preserve the numerical retained-diversity ratio: if $H'_{ij} = H_{ij}/c_i$ for a positive thread-level constant $c_i$, then

$$R'_{D,i} = 2^{(H_{i,\mathrm{late}} - H_{i,\mathrm{early}})/c_i} = R_{D,i}^{1/c_i}.$$

The weak empirical association of $R_{D,i}$ with schema size is separate from rank invariance. Neither this check nor capacity normalisation establishes that different schemas represent equal semantic breadth or recognise the same evaluations.

# Supplementary Tables

Tables S1–S5 give the full estimates behind Section 4 of the main text: the four classifications and the magnitude of narrowing (Table S1), narrowing by thread size (Table S2), the opening-composition analysis (Table S3), the sensitivity of the contraction-episode results to the alternative episode definitions (Table S4) and the April 2018 estimates under the three relevance-gate rules (Table S5). Table S6 reports the human audit of the model-generated labels (Supplementary Methods S8 and Results R2). Intervals in these tables are 95% community-cluster bootstrap percentile intervals unless an entry is explicitly an interquartile range. Regression intervals in Figure S5 use clustered standard errors.

**Table S1 Whole-thread tunnel vision classifications and narrowing.** Whole-thread narrowing in the analysis sample (2,967 threads): the four classifications of tunnel vision and the magnitude of narrowing in full detail. Shares and medians carry 95% community-cluster bootstrap intervals; interquartile ranges are over threads. Observed tunnel vision: $\rho_{\mathrm{ASPE}} < -0.15$ and $\rho_{\mathrm{conc}} > 0$; order-robust: order-shuffle $p < 0.05$; density-robust: density-preserving $p < 0.05$, the subset emphasised in the dynamic analyses; dual-null-robust: both. Each robustness descriptor additionally requires observed tunnel vision. The two nulls test different null hypotheses and the order-robust and density-robust subsets are not nested: among the 1,363 threads with observed tunnel vision, 179 are dual-null-robust, 228 order-robust only, 97 density-robust only and 859 neither (Supplementary Figure S1); the remaining 1,604 threads are non-tunnel-vision threads. "All other threads" are the analysis-sample threads that are not density-robust.

| | All threads | Density-robust tunnel vision | Dual-null-robust tunnel vision | All other threads |
|---|---|---|---|---|
| *Classifications (share of the analysis sample)* | | | | |
| Observed tunnel vision | 1,363 (45.9%; 43.6–48.1) | – | – | – |
| Order-robust tunnel vision | 407 (13.7%; 12.5–15.0) | – | – | – |
| Density-robust tunnel vision (primary) | 276 (9.3%; 7.8–10.6) | – | – | – |
| Dual-null-robust tunnel vision | 179 (6.0%; 5.0–6.9) | – | – | – |
| *Magnitude* | | | | |
| Threads | 2,967 | 276 | 179 | 2,691 |
| Median $\rho_{\mathrm{ASPE}}$ (95% interval) | -0.12 (-0.14, -0.10) | -0.53 (-0.56, -0.50) | -0.60 (-0.64, -0.57) | -0.09 (-0.10, -0.07) |
| interquartile range | -0.34 to 0.10 | -0.64 to -0.42 | -0.70 to -0.53 | -0.28 to 0.12 |
| Median $\Delta H$ (bits) (95% interval) | -0.12 (-0.15, -0.10) | -0.43 (-0.49, -0.38) | -0.50 (-0.55, -0.43) | -0.09 (-0.12, -0.07) |
| interquartile range | -0.45 to 0.18 | -0.73 to -0.12 | -0.81 to -0.21 | -0.41 to 0.21 |
| Median $R_D$ (95% interval) | 0.92 (0.90, 0.93) | 0.74 (0.71, 0.77) | 0.71 (0.68, 0.74) | 0.94 (0.92, 0.95) |
| interquartile range | 0.73 to 1.13 | 0.60 to 0.92 | 0.57 to 0.87 | 0.75 to 1.16 |
| $\rho_{\mathrm{ASPE}} < 0$ | 64% (1,907/2,967) | 100% (276/276) | 100% (179/179) | 61% (1,631/2,691) |
| $\rho_{\mathrm{ASPE}} \leq -0.5$ | 10% (311/2,967) | 57% (158/276) | 84% (150/179) | 6% (153/2,691) |
| $\Delta H < 0$ | 62% (1,826/2,967) | 86% (237/276) | 90% (161/179) | 59% (1,589/2,691) |
| Density-robust tunnel vision minus all other threads, difference of median $R_D$: -0.19 (-0.22, -0.16) | | | | |

**Table S2 Density-robust tunnel vision by thread length.** Density-robust tunnel vision by usable-size band within the predefined analysis frame: no clear monotonic prevalence pattern is apparent across the usable-size bands (top; 95% community-cluster bootstrap intervals; the two coarser bands used in the analysis are repeated below the fine bins) and the 14 supplementary long threads whose usable count exceeds 600 (bottom; within the platform-reported comment band at archiving but above the usable-size range of the analysis sample, listed for completeness and not used for inference).

| Usable comments | $n$ | Density-robust tunnel vision | Mean $\rho_{\mathrm{ASPE}}$ |
|---|---|---|---|
| 150 to 199 | 36 | 11.1% (2.5–22.2) | −0.20 (−0.33, −0.09) |
| 200 to 299 | 416 | 9.4% (7.0–12.4) | −0.13 (−0.16, −0.11) |
| 300 to 399 | 1,363 | 10.3% (8.4–12.0) | −0.13 (−0.15, −0.12) |
| 400 to 499 | 815 | 8.1% (5.6–10.2) | −0.09 (−0.11, −0.06) |
| 500 to 600 | 337 | 8.0% (5.4–10.9) | −0.05 (−0.09, −0.02) |
| 150 to 299 | 452 | 9.5% (7.2–12.5) | −0.14 (−0.17, −0.11) |
| 300 to 600 | 2,515 | 9.3% (7.6–10.8) | −0.11 (−0.12, −0.09) |

*Supplementary long threads (usable > 600): community: usable comments; $\rho = \rho_{\mathrm{ASPE}}$; observed tunnel vision; $p_{\mathrm{order}}$; $p_{\mathrm{density}}$; $R_D$*
r/Futurology: 3,239; $\rho = -0.00$; observed no; $p_{\mathrm{order}} = 0.48$; $p_{\mathrm{density}} = 0.65$; $R_D = 0.76$
r/worldpolitics: 1,846; $\rho = 0.13$; observed no; $p_{\mathrm{order}} = 0.88$; $p_{\mathrm{density}} = 0.63$; $R_D = 0.73$
r/worldnews: 1,555; $\rho = -0.00$; observed no; $p_{\mathrm{order}} = 0.48$; $p_{\mathrm{density}} = 0.57$; $R_D = 1.13$
r/worldnews: 1,467; $\rho = 0.00$; observed no; $p_{\mathrm{order}} = 0.46$; $p_{\mathrm{density}} = 0.63$; $R_D = 0.88$
r/nottheonion: 1,343; $\rho = -0.01$; observed no; $p_{\mathrm{order}} = 0.51$; $p_{\mathrm{density}} = 0.21$; $R_D = 1.47$
r/nottheonion: 1,323; $\rho = -0.30$; observed yes; $p_{\mathrm{order}} = 0.02$; $p_{\mathrm{density}} = 0.42$; $R_D = 0.87$
r/atheism: 1,316; $\rho = -0.28$; observed yes; $p_{\mathrm{order}} = 0.03$; $p_{\mathrm{density}} = 0.30$; $R_D = 1.02$
r/Futurology: 1,235; $\rho = -0.36$; observed yes; $p_{\mathrm{order}} = 0.02$; $p_{\mathrm{density}} = 0.12$; $R_D = 1.43$
r/Futurology: 862; $\rho = -0.18$; observed yes; $p_{\mathrm{order}} = 0.16$; $p_{\mathrm{density}} = 0.17$; $R_D = 0.47$
r/LateStageCapitalism: 819; $\rho = -0.05$; observed no; $p_{\mathrm{order}} = 0.41$; $p_{\mathrm{density}} = 0.12$; $R_D = 0.69$
r/UpliftingNews: 657; $\rho = -0.32$; observed yes; $p_{\mathrm{order}} = 0.05$; $p_{\mathrm{density}} = 0.04$; $R_D = 0.45$
r/UpliftingNews: 607; $\rho = -0.28$; observed yes; $p_{\mathrm{order}} = 0.08$; $p_{\mathrm{density}} = 0.14$; $R_D = 0.97$
r/relationship_advice: 605; $\rho = 0.22$; observed no; $p_{\mathrm{order}} = 0.82$; $p_{\mathrm{density}} = 0.81$; $R_D = 0.97$
r/ukpolitics: 602; $\rho = -0.25$; observed yes; $p_{\mathrm{order}} = 0.11$; $p_{\mathrm{density}} = 0.53$; $R_D = 1.38$

**Table S3 Subsequent diversity by opening-ASPE quartile.** Opening composition and what follows it: subsequent narrowing by quartile of opening ASPE (2,967 analysis-sample threads, equal weight per thread). Entries are medians with interquartile ranges unless stated; the opening-ASPE column gives the median and the quartile's range; shares of density-robust tunnel vision carry 95% community-cluster bootstrap intervals; other percentage entries are descriptive point estimates. $\rho_{H,\mathrm{post}}$ and $\Delta H_{\mathrm{post}}$ use only windows after comment 50; $H_{\mathrm{late}}$ is the ASPE of the last full window.

| Quartile | $n$ | Opening ASPE | $\rho_{H,\mathrm{post}}$ | $\rho_{H,\mathrm{post}} < 0$ | $\Delta H_{\mathrm{post}}$ (bits) | $H_{\mathrm{late}}$ (bits) | Density-robust tunnel vision | Dual-null-robust tunnel vision |
|---|---|---|---|---|---|---|---|---|
| Q1 (lowest) | 742 | 2.93 (1.24–3.17) | −0.07 (−0.30, 0.17) | 57% | −0.10 (−0.50, 0.24) | 2.99 (2.65, 3.33) | 4.0% (2.7–5.5) | 2.0% |
| Q2 | 742 | 3.32 (3.17–3.46) | −0.13 (−0.34, 0.12) | 63% | −0.13 (−0.45, 0.23) | 3.25 (2.94, 3.53) | 9.2% (6.7–11.6) | 5.0% |
| Q3 | 741 | 3.61 (3.46–3.76) | −0.11 (−0.31, 0.14) | 62% | −0.12 (−0.40, 0.21) | 3.45 (3.17, 3.69) | 8.6% (5.4–11.1) | 5.7% |
| Q4 (highest) | 742 | 3.95 (3.76–4.74) | −0.14 (−0.37, 0.09) | 67% | −0.15 (−0.42, 0.11) | 3.71 (3.42, 3.94) | 15.4% (11.9–18.4) | 11.5% |

**Table S4 Sensitivity to alternative episode definitions.** Sensitivity of the local contraction results to the pre-specified alternative episode definitions (2,967-thread analysis sample; Supplementary Figure S7 shows the same settings with intervals). The whole-thread measurement window remains fixed at $W = 30$. T1 and T2 change only the onset/recovery thresholds; R0 and R1 begin episode detection after comment 50, using the first-window reference (R0) or the equivalent diversity implied by the mean entropy of the five opening windows (R1). Columns report the share of threads with an episode, the share of first episodes that recover, the share persisting to the end among threads with at least 30 comments of follow-up, and the median relative position of the first onset.

| Setting | Rule | Episode (%) | Recovered (%) | Persisted (%) | Onset (relative) |
|---|---|---|---|---|---|
| Main | 0.80/0.90, all windows | 73.8 | 83.7 | 15.1 | 0.28 |
| T1 | 0.75/0.85, all windows | 65.4 | 84.7 | 10.1 | 0.32 |
| T2 | 0.85/0.95, all windows | 80.6 | 80.7 | 21.0 | 0.23 |
| R0 | 0.80/0.90, after comment 50, first-window reference | 72.1 | 80.2 | 16.9 | 0.35 |
| R1 | 0.80/0.90, after comment 50, opening-mean reference | 73.5 | 82.5 | 11.9 | 0.37 |

**Table S5 April 2018 sensitivity to relevance-gate rules.** April 2018 estimates under three relevance-gate rules, on the same 1,035 threads and the same frozen schemas: the rule of the 2019 census driver (used in the main text), the rule as documented in the earlier study source (regex name tokens longer than three characters matched as words, cues as substrings) and the rule of the local classifier service (name tokens longer than two characters minus stop-words, single-word cues as words). Entries are thread-equal-weight estimates with 95% community-cluster bootstrap intervals. The classifier, schemas, thread files and null seeds are identical across columns. Recovery is conditional on a first episode, relapse on its recovery, and persistence on an episode with sufficient follow-up.

| | census rule (primary) | documented rule | service rule |
|---|---|---|---|
| Observed tunnel vision (%) | 38.9 (35.7, 41.7) | 41.1 (37.5, 44.0) | 40.8 (37.3, 43.2) |
| Density-robust tunnel vision (%) | 7.8 (6.4, 9.4) | 7.0 (5.2, 8.7) | 6.4 (4.9, 7.8) |
| Dual-null-robust tunnel vision (%) | 3.9 (3.0, 4.8) | 3.2 (2.3, 4.1) | 2.8 (1.9, 3.8) |
| Median $R_D$, all threads | 0.938 (0.906, 0.961) | 0.946 (0.924, 0.975) | 0.953 (0.925, 0.975) |
| Median $R_D$, density-robust tunnel vision | 0.744 (0.706, 0.815) | 0.797 (0.680, 0.839) | 0.726 (0.672, 0.816) |
| Threads with an episode (%) | 71.1 (68.3, 73.5) | 72.0 (68.5, 74.9) | 73.3 (70.0, 76.3) |
| First episodes recovered (%) | 81.5 (79.3, 84.0) | 83.6 (81.5, 86.1) | 84.1 (81.6, 87.4) |
| Persisted to the end (%) | 16.9 (14.4, 19.0) | 16.0 (13.7, 18.8) | 14.7 (12.3, 16.8) |
| Relapse after first recovery (%) | 71.3 (67.8, 75.4) | 69.8 (66.2, 72.7) | 70.1 (66.3, 73.0) |
| Returning share, paired change at recovery (pp) | 12.5 (10.7, 13.9) | 10.3 (8.6, 11.6) | 11.1 (9.6, 12.6) |
| First-observed share, paired change at recovery (pp) | -0.1 (-1.4, 1.5) | 1.3 (0.3, 2.3) | 0.5 (-0.5, 1.5) |
| $\Delta H$ within the common set at recovery (bits) | 0.036 (0.022, 0.055) | 0.031 (0.020, 0.046) | 0.034 (0.024, 0.047) |
| Previous dominant pair, share change at recovery (pp) | -17.4 (-18.8, -15.7) | -14.7 (-16.2, -12.7) | -15.3 (-16.6, -13.5) |
| Median coverage (share of comments with a label) | 0.455 | 0.470 | 0.450 |
| Median mentions per comment | 0.74 | 0.81 | 0.75 |

**Table S6 Human annotation agreement and pipeline performance.** Human audit of the model-generated labels (Supplementary Methods S8): 501 comments from 42 threads labelled independently by two annotators. Agreement is reported at three levels, detection (whether an aspect is addressed), polarity (where both sources record a pair) and position in the thread, followed by comment-level and thread-level summaries. Pipeline estimates other than the per-annotator rows are computed within the consensus cells, the cells on which both annotators assigned the same label (disputed cells are excluded), with 95% thread-stratified bootstrap intervals; inter-annotator statistics use all cells. Detection recall is the share of consensus pairs for which the pipeline records a pair of any polarity, and detection precision the share of its pairs in those cells that are consensus pairs; polarity agreement is conditional on detection; exact-pair recall and precision require both aspect and polarity to match.

| | Estimate (95% interval) |
|---|---|
| *Level 1: detection (whether an aspect is addressed)* | |
| Annotators: aspect addressed or not, cell agreement ($\kappa$) | 91.3% (0.48) |
| Pipeline against annotator A / B: cell agreement ($\kappa$) | 93.9% (0.62) / 90.7% (0.41) |
| Pipeline within the consensus cells (216 consensus pairs): detection recall | 0.78 (0.70, 0.85) |
| detection precision | 0.56 (0.45, 0.68) |
| Detection recall, density-robust / other threads | 0.83 (0.69, 0.93) / 0.75 (0.64, 0.84) |
| Detection precision, density-robust / other threads | 0.46 (0.32, 0.62) / 0.67 (0.54, 0.80) |
| Detection recall / precision, April 2018 threads ($n = 96$ comments) | 0.93 (0.87, 1.00) / 0.93 (0.87, 1.00) |
| *Level 2: polarity (pairs recorded by both sources)* | |
| Annotators: polarity agreement where both record a pair, all cells ($n = 290$) | 74.5% |
| Pipeline: polarity agreement on detected consensus pairs ($n = 169$) | 95.9% |
| Exact-pair recall (aspect and polarity), 216 consensus pairs | 0.75 (0.66, 0.82) |
| Exact-pair precision | 0.54 (0.43, 0.65) |
| Exact-pair recall, April 2018 threads | 0.91 (0.84, 0.97) |
| *Level 3: position in the thread (pipeline within the consensus cells)* | |
| Detection recall, early / middle / late third | 0.74 (0.59, 0.86) / 0.79 (0.67, 0.88) / 0.81 (0.69, 0.91) |
| Detection recall, early minus late | −0.08 (−0.22, +0.06) |
| Exact-pair recall, early / middle / late third | 0.68 (0.54, 0.80) / 0.78 (0.64, 0.88) / 0.79 (0.64, 0.90) |
| Exact-pair recall, early minus late | −0.11 (−0.26, +0.05) |
| Detection precision, early / middle / late third | 0.55 (0.42, 0.69) / 0.51 (0.37, 0.67) / 0.62 (0.49, 0.74) |
| Detection precision, early minus late | −0.07 (−0.18, +0.05) |
| Detection precision, density-robust threads, early / late third | 0.46 (0.26, 0.64) / 0.47 (0.32, 0.65) |
| *Comment-level summaries (pair sets per comment)* | |
| Annotators: mean per-comment pair-set F1 | 0.39 |
| comments with identical pair sets | 27.7% |
| mean F1, comments flagged as sarcastic ($n = 23$) / all others | 0.17 / 0.40 |
| Pipeline against annotator A / B: mean per-comment pair-set F1 | 0.65 / 0.32 |
| pair set identical to at least one annotator | 60.5% |
| *Thread level (42 threads, 11–12 comments each)* | |
| Distinct pairs, Spearman: pipeline–annotator A / pipeline–annotator B / A–B | 0.76 / 0.55 / 0.28 |
| Share of negative pairs, early → late third: annotator A / B / pipeline | 54%→65% / 47%→63% / 52%→75% |

# Supplementary Figures

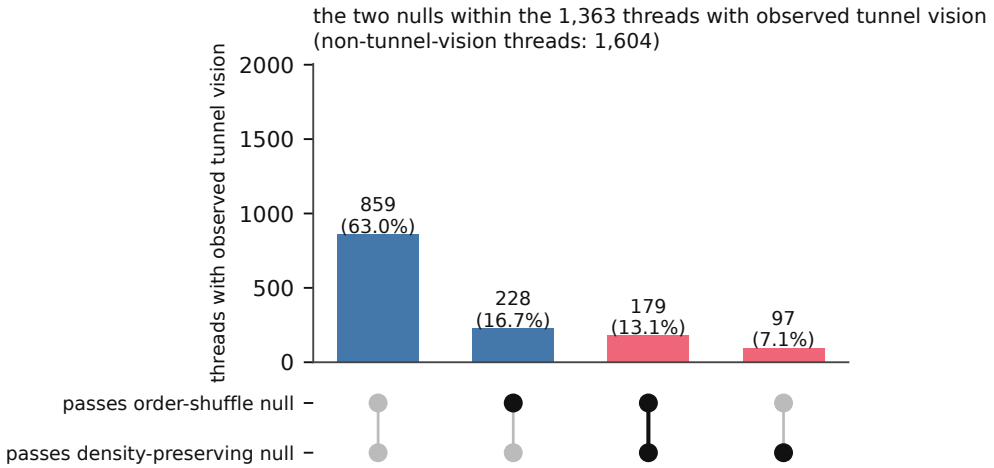



**Fig. S1 Overlap of the two null-based classifications.** The two null comparisons within the 1,363 threads with observed tunnel vision (analysis sample). Bars count the exclusive combinations: neither null passed, order-shuffle only, both (dual-null-robust), density-preserving only (density-robust but not order-robust); the dot matrix below marks which null is passed. The 1,604 non-tunnel-vision threads are outside the figure. The two nulls test different alternative explanations and the order-robust and density-robust subsets are not nested.

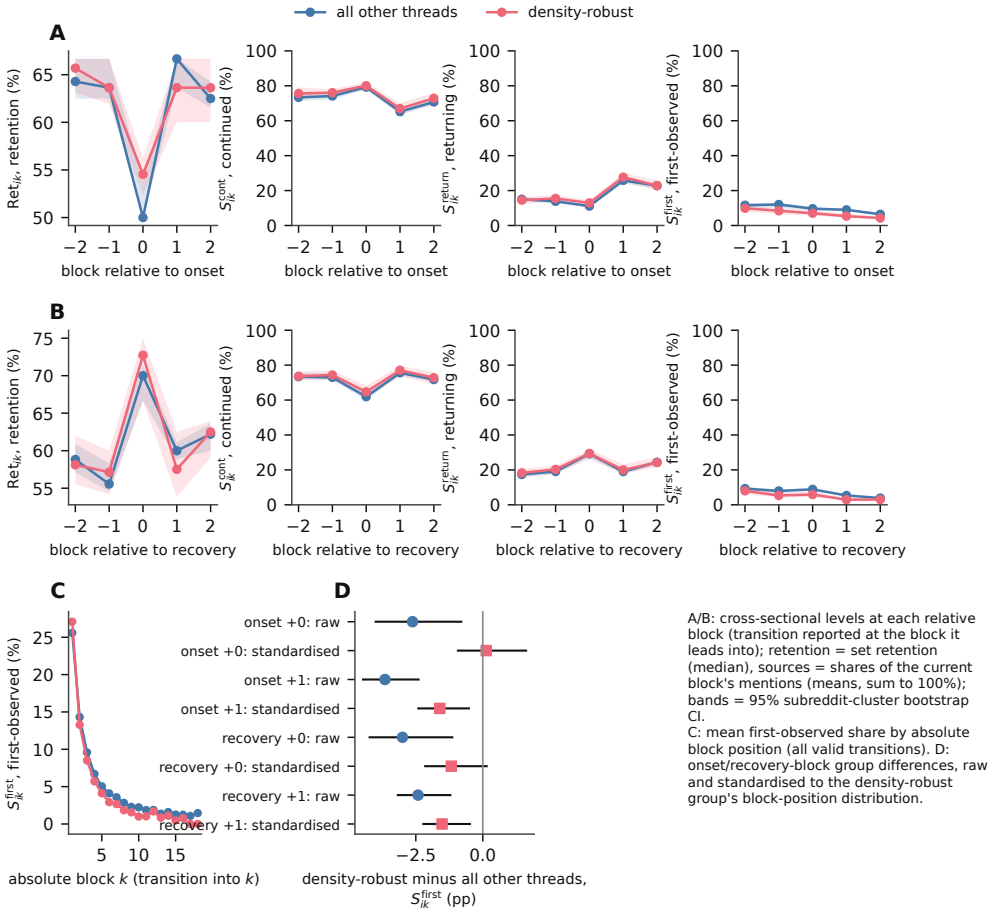



**Fig. S2 Pair retention and mention sources around first episodes.** Block-level detail around the first episode, threads with density-robust tunnel vision (red) against all other threads (blue). (a, b) Cross-sectional levels at each block relative to the onset (a) and the recovery (b) block: retention $\mathrm{Ret}_{ik}$ (median) and the shares $S_{ik}^{\mathrm{cont}}$, $S_{ik}^{\mathrm{return}}$ and $S_{ik}^{\mathrm{first}}$ of the block's mentions (means); bands are 95% community-cluster bootstrap intervals. (c) Mean first-observed share by absolute block position $k$, all valid transitions: the share falls steeply with position in both groups. (d) Difference between density-robust and all other threads in the first-observed share at the onset and recovery blocks and one block later, raw (circles) and standardised to the density-robust group's block-position distribution (squares), with 95% cluster intervals.

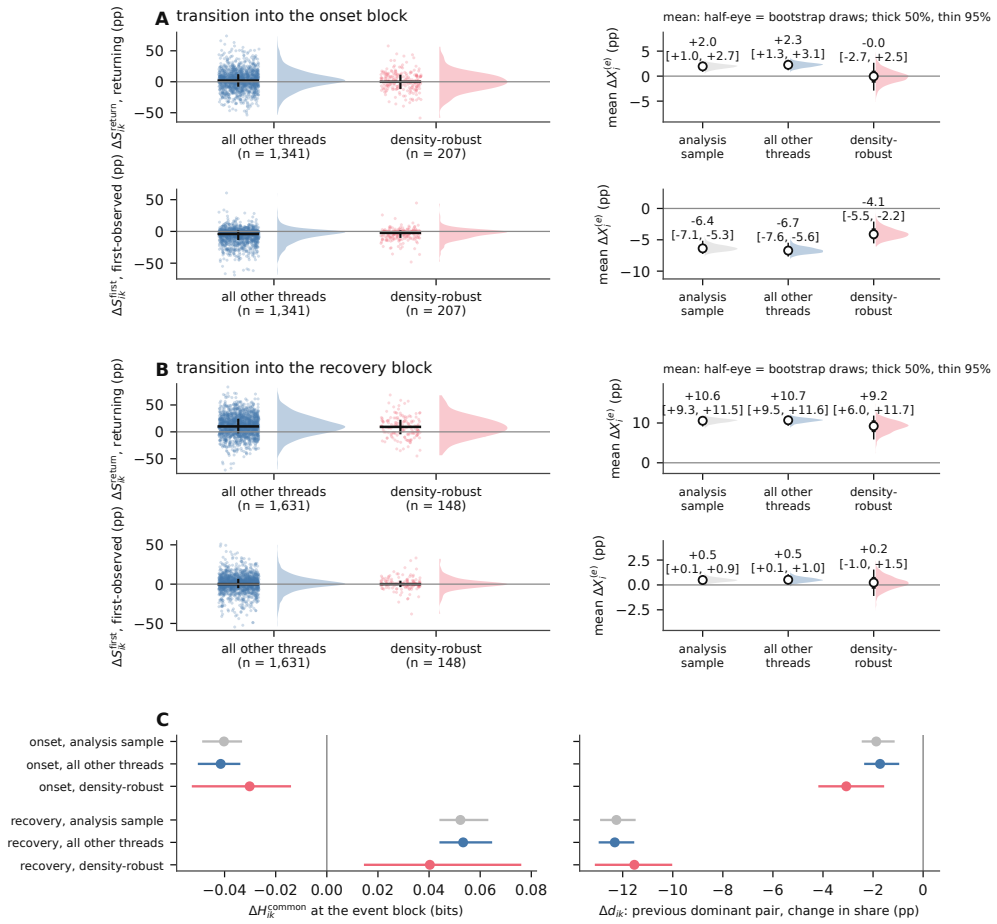



**Fig. S3 Paired changes at onset and recovery.** Paired changes into the onset block (a) and the recovery block (b) by group. Left: per-thread paired change $\Delta X_i^{(e)}$ of the returning share $S_{ik}^{\text{return}}$ and the first-observed share $S_{ik}^{\text{first}}$ for all other threads and threads with density-robust tunnel vision (points, half-violin, median and interquartile range). Right: mean paired change for the analysis sample, all other threads and threads with density-robust tunnel vision, with the bootstrap draws as a half-eye and the 50% (thick) and 95% (thin) community-cluster intervals. (c) Two already-differenced quantities at the event block by group: the entropy change within the common pair set $\Delta H_{ik}^{\text{common}}$ (bits) and the change in share of the previous block's dominant pair $\Delta d_{ik}$ (percentage points), means with 95% cluster intervals.

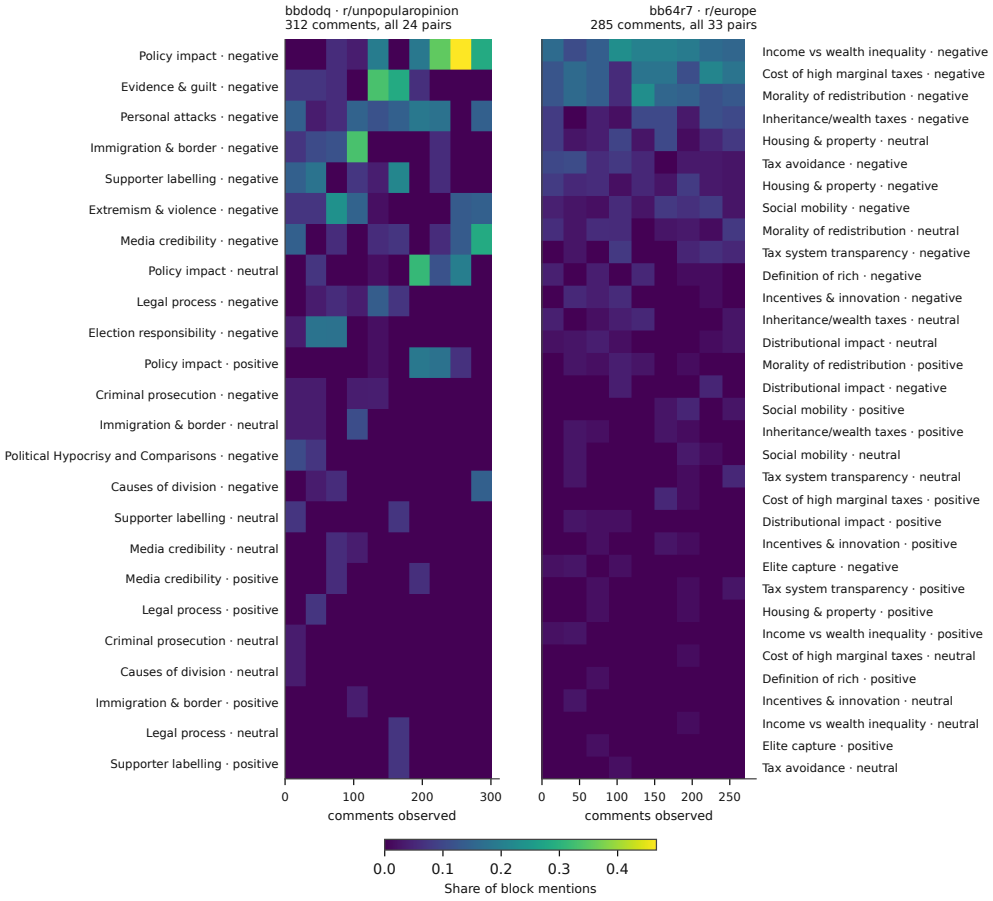



**Fig. S4 Complete pair trajectories for two illustrative threads.** The two illustrative threads with density-robust tunnel vision of main-text Figure 5 with every observed aspect-sentiment pair: share of each 30-comment block's mentions per pair, same colour scale as main-text Figure 5; hatched columns are blocks without a retained mention.

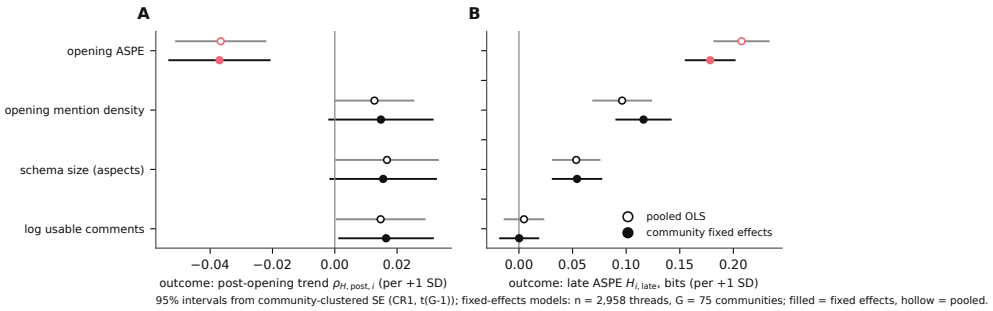
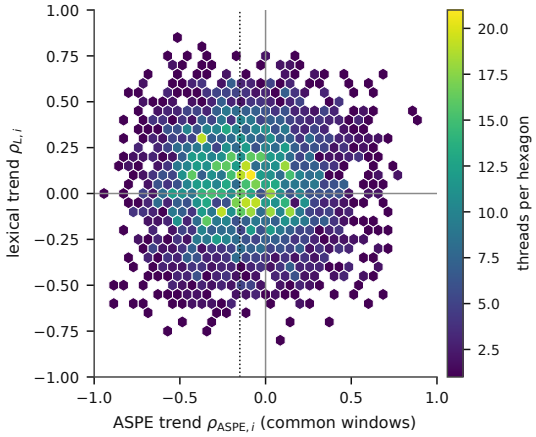
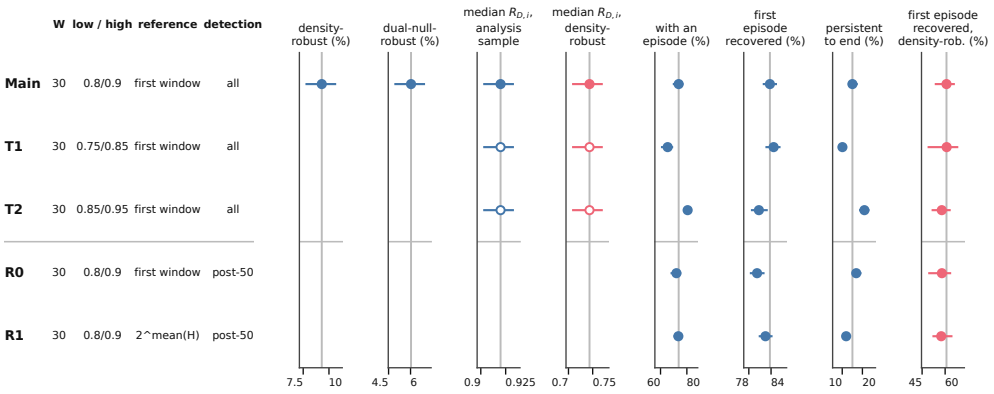



**Fig. S5 Adjusted associations of opening and later diversity.** Coefficients of the adjusted linear models of Supplementary Methods S5, per one standard deviation of each covariate, for the post-opening trend $\rho_{H,\text{post},i}$ (a) and the late ASPE $H_{i,\text{late}}$ (b): opening ASPE (red) and the three control variables, under the pooled specification (hollow) and with community fixed effects (filled); bars are 95% intervals from community-clustered standard errors.



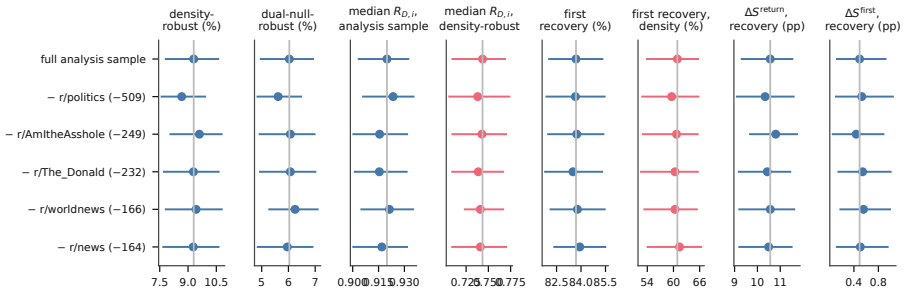

Spearman correlation of the two trends
analysis sample (n = 2,967): +0.002 [-0.037, +0.040]
density-robust (n = 276): -0.122 [-0.260, +0.027]
95% community-cluster bootstrap intervals; dotted line: threshold of observed tunnel vision (−0.15)

**Fig. S6 Evaluative and lexical diversity trends.** Evaluative and lexical diversity trends are largely unrelated. Thread-level trend $\rho_{L,i}$ of lexical diversity (distinct word types per 100 sampled tokens; Supplementary Methods S5) against the ASPE trend $\rho_{\mathrm{ASPE},i}$ on the same windows, as a hexagonal density of the 2,967 analysis-sample threads; the dotted line marks the threshold −0.15 of observed tunnel vision on the ASPE axis, and the Spearman correlations beneath the panel carry 95% community-cluster bootstrap intervals.




points = estimate with 95% subreddit-cluster bootstrap CI; hollow = inherited from the main setting (not recomputed). grey vertical line = main setting. T1/T2 change only the onset/recovery thresholds; R0/R1 detect episodes only after comment 50 and are comparable only with each other (common threads).

**Fig. S7 Episode-definition sensitivity with uncertainty intervals.** Sensitivity of the main estimates to the pre-specified alternative episode definitions (Supplementary Table S4): the main setting, the two alternative onset/recovery thresholds (T1, T2) and the two post-opening detection rules (R0, R1). Points are thread-equal-weight estimates with 95% community-cluster bootstrap intervals; hollow points are inherited from the main setting and not recomputed; the grey vertical line marks the main setting. The whole-thread measurement window $W = 30$ is fixed and not varied.




each row drops one of the five largest communities (by analysis-sample threads); estimates with 95% cluster-bootstrap CI over the remaining communities; grey line = full sample.

**Fig. S8 Leave-one-community-out robustness.** Influence of the five largest communities: each row re-aggregates the frozen per-thread results after dropping one community, with 95% community-cluster bootstrap intervals over the remaining communities; the grey vertical line marks the full analysis sample.

# Terminology used throughout the supplement

The supplement follows the same terminology as the main text. *Observed tunnel vision* denotes a thread satisfying the pre-specified whole-thread trend criterion. *Density-robust tunnel vision*, *order-robust tunnel vision* and *dual-null-robust tunnel vision* denote robustness of observed tunnel vision to the corresponding specified alternatives. *Non-tunnel-vision threads* do not meet the whole-thread trend criterion. When threads with density-robust tunnel vision are compared with the remainder of the analysis sample, the comparison group is *all other threads*. Local *contraction episodes*, *onset*, *recovery*, *persistence* and *relapse* are trajectory descriptors and are not whole-thread tunnel-vision classifications. An evaluative position is an aspect-sentiment pair; a pair mention is an occurrence of that pair in the retained label stream. Continued, returning and first-observed pairs are the mutually exclusive block-transition categories defined in the pair-level block analysis. Equivalent diversity is the effective number of equally weighted pairs; relative diversity is equivalent diversity divided by its reference value. Recovery refers to the relative-diversity threshold, while return refers to the reappearance of a pair.

# References

[1] Oprea, S., Magdy, W.: isarcasm: A dataset of intended sarcasm. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1279–1289 (2020)

[2] Plank, B.: The "problem" of human label variation: On ground truth in data, modeling and evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi (2022). https://aclanthology.org/2022.emnlp-main.731/

[3] Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., Jovančević, A.: Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. Scientific Reports **15**, 11477 (2025) https://doi.org/10.1038/s41598-025-96508-3