Differentiation Drives the Erosion of Positivity on Social Media

Hongkai Mao[1,2*], Yuan Chang Leong[3,4], Yutong Jiang[1], Alex Koch[1], William J Brady[5], Joshua Conrad Jackson[1,4*]

1. Booth School of Business, University of Chicago
2. Graduate School of Business, Stanford University
3. University of Chicago, Department of Psychology
4. Data Science Institute, University of Chicago
5. Kellogg School of Management, Northwestern University

* Corresponding authors:

Hongkai Mao
hm404@stanford.edu

Joshua Conrad Jackson
joshua.jackson@chicagobooth.edu

Keywords: Social Media; Computational Social Science; Cultural Evolution; Information Ecologies; Social Psychology

Abstract (200/250 Words)

Most people believe that social media discourse is negative and divisive. Here we show how this negativity can evolve even when users are not motivated to be negative. We propose that social media users seek to differentiate themselves from other users, and it is easier to differentiate oneself through negativity than positivity because negative information is more heterogeneous and counter-normative than positive information. This makes users increasingly likely to post negative comments as a conversation unfolds and it becomes more challenging to make unique contributions. Analyzing 2.05 billion comments from 2,150 Reddit communities shows that comments become more negative over time, both within threads and community histories. This trend towards negativity is mediated by the semantic uniqueness of comments, suggesting that it arises from users differentiating themselves. This trend is strongest when initial dialogue is positive, making negative comments highly counter-normative. We replicate these patterns in a multigenerational experiment simulating social media dialogue (*n* = 4,000). Participants become more negative over time, but only when incentivized to be unique, and especially when dialogue begins positively. These findings suggest that the structure of social media platforms interacts with human motivation to foster a drift towards negativity over time in online discourse.

Significance Statement (119/120 Words)

We live in a digital age, where billions of people engage in dialogue within topic-bound communities and threads. In an archival analysis of over 2 billion Reddit comments and a multi-generational experiment, we show that this dialogue becomes more negative over time. Further analyses suggest that negativity rises over time because social media users seek to make unique comments on the same topic, and it is easier to differentiate oneself through negative comments than through positive comments. As threads and communities evolve, and it becomes more difficult to make unique observations, users turn to negativity. Our studies show how basic human motives interact with the structure of social media platforms, posing an acute challenge for sustaining healthy online dialogue.

Differentiation Drives the Erosion of Positivity On Social Media

One of the earliest theories of online discourse was Godwin's law–the adage that, as an online discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches one. While Godwin's law does not appear to be literally accurate (1), it also raises a broader prediction: that online conversations tend to devolve into negativity over time. Here, we examine this prediction by analyzing whether social media discourse becomes more negative as discussions evolve, and testing why this happens.

Understanding the roots of negativity on social media is a major goal for both science and society (2–6). Most social media users believe that online platforms are divisive, toxic, and have a harmful impact on society (7,8). One leading explanation of this negativity is that social media incentivizes negative content because it elicits more attention and engagement than positive content (5,7,9). Engagement-based algorithms therefore amplify negative posts, pushing them to the top of users' feeds (3,10,11). This visibility may lead users to internalize norms of negativity or strategically post negative content to attract attention themselves. In this model, users who post negative content seek to do so–either to conform to platform norms or to attract engagement.

Here we explore why users may also post negative content as a byproduct of a more benevolent motivation: the desire to differentiate oneself. Psychologists have long recognized differentiation as a fundamental human motive (12), which influences group behavior in offline contexts (13). However, social media accentuates this motivation by rewarding unique posts with shares, likes, and views (7). Multiple studies have now found that novelty is associated with online virality (14,15), that using social media is linked to rises in the need for uniqueness (16), and that users actively seek to differentiate themselves in their posting behavior over time (17). Social media platforms are therefore called "attention economies" in which users seek to stand out from the crowd (18). Ironically, however, the structure of social media also makes it hard for users to stand out. Social media threads involve sustained discussion of a single topic. As time passes and more people comment, users may find it increasingly challenging to make unique contributions to these discussions and draw attention to their comments.

We propose that, as users find it increasingly difficult to stand out, they may resort to negativity as a means of differentiating themselves. Negativity facilitates differentiation for two reasons. First, negative information is more semantically heterogeneous than positive information, echoing Tolstoy's famous adage that: "all happy families are alike; each unhappy family is unhappy in its own way." Many studies of "information ecology" (19,20) now support Tolstoy's claim, showing that there are more diverse ways of communicating negative (versus positive) information about emotion (21,22) and describing other people (23,24). Second, negative information is more counter-normative than positive information because positive comments are more common than negative ones (25), which makes negative comments stand out (26). Because negativity is heterogeneous and counter-normative, users may unintentionally write negative comments from a desire to stand out in crowded online spaces.

Under this explanation, we would expect to see four patterns in social media data. First, comments should not be immediately negative. Instead, comments should become negative over time as users find it more difficult to differentiate themselves. This trend should characterize social media spaces where users are topic-constrained, like threads or communities devoted to a specific subject. In contrast, it should not characterize other spaces like online news, which do not have this sequential comment structure.

Second, semantic differentiation should statistically explain rises in negativity. Negativity should increase most reliably when it allows users to differentiate themselves from their peers. Likewise, there should only see rises in negativity when users seek to be distinct. If users instead seek to conform, the trend may disappear or reverse.

Third, this trend toward negativity should be strongest when conversations begin positively. In these cases, negative comments are not only more varied but also more counter-normative–making them more distinctive in two ways. In contrast, when conversations begin negatively, both positive and negative comments may stand out in different ways (negativity is still more heterogeneous, but positivity is more counter-normative), so the trend towards negativity should be weaker or non-existent because positive and negative comments should *both* rise over time. These same dynamics should also lead trends towards negativity to flatten over time, as they reach an equilibrium where positive and negative comments can both stand out.

Fourth, the number of people in a conversation should mirror the effect of time. Larger communities should be more negative because the challenge of standing out intensifies when more people are competing to speak on the same topic. In other words, rises in the size of social media communities should predict greater negativity above and beyond the passage of time.

Current Research

These four patterns constitute four hypotheses that we test using a mix of observational and experimental data. Our observational analysis draws from 2.05 billion comments from 2,150 topic-specific “subreddit” communities on Reddit across 2010 - 2022. Reddit’s scale and topical structure make it ideal for testing our hypotheses. We test whether threads and subreddits become more negative over time, whether these trends are mediated by semantic differentiation, whether they are strongest when initial dialogue is positive, and whether larger communities are more negative than smaller communities. To benchmark our findings, we compare these results with 4.91 million news articles from 267 publishers posted over the same period. Unlike Reddit, these news articles do not show increasing negativity over time.

We then run a large-scale experiment in which 4,000 participants post comments across five “generations” of an evolving social media feed. We manipulate the motivation to be unique by either incentivizing participants to either contribute unique content or to conform to prior content (i.e., disincentivizing the motive to be unique). In the uniqueness incentive condition, we find a pattern that closely mirrors our observational findings: participants contribute more negative content over time, especially when dialogue starts out positively. In the conformity incentive condition, we find a different pattern that does not resemble our observational data. This experiment conceptually replicates our archival analyses and allows for causal inference about the role of social incentives.

All regression specifications, methods for tracking semantic uniqueness, and robustness checks are described in detail in the Methods and Supplementary Materials. Coefficients are standardized to facilitate interpretation, and all data and code are available at https://osf.io/68wvm/.

Results

Observational Evidence From 2 Billion Reddit Comments

Increased Negativity Over Time. We first hypothesized that threads should become more negative over time both at the thread and community level. In a series of mixed effects regressions, we found support for both of these predictions.

We first analyzed a sample of 115,961 threads (5.5 million comments), randomly drawn from all threads with at least 50 comments. We truncated the size of the thread at 50, so that we did not confound changes as threads became longer with cohort effects (inherent differences between the topics that inspire longer threads and shorter threads). Our regression also nested comments within their threads and communities to adjust standard errors, and controlled for the comment level (i.e., whether a comment was a direct reply to the original post or a reply to another comment). In this model, we found that comments posted later in threads were associated with more negative valence, $b = -0.003$, $SE = 0.0005$, $t = -5.54$, $p < 0.001$, *95% CIs* [-0.004, -0.002].

Threads could become more negative over time because of two processes. The first involves comment proportions: users might post more negative comments, fewer positive comments, or both. The second involves comment intensity: negative comments might become more extremely negative, positive comments less extremely positive, or both. We find evidence for both processes. As shown in **Fig. 1a**, the proportion of negative comments increased over the course of threads. Among the first 5 comments in each thread, 25.21% were classified as negative, compared to 25.96% among the final 5 comments—a modest increase of 0.75%. Although this difference is small, it is highly robust. When aggregating comments by their sequential position, the proportion of negative comments strongly increased over time ($R^2 = 0.69$).

In addition, negative comments became more intensely negative (**Fig. 1a**). The average valence of negative comments declined from -0.490 among the first 5 comments to -0.498 among the last 5. Interestingly, as shown in **Fig. 1b,** the proportion of positive comments also increased slightly (from 50.85% to 51.11%), but the intensity of positive comments remained stable, with virtually no change in their average valence (< 0.01%).

We next tested our hypothesis at the community level, examining whether entire communities became more negative over time. We sampled the average monthly valence of comments for each community, and then regressed monthly valence against the passage of time while nesting observations in communities and month to adjust our standard errors. In this analysis, we again found that time was associated with a decline in valence, *b* = -0.03, *SE* = 0.003, *t* = -9.91, *p* < 0.001, *95% CIs* [-0.04, -0.02]. Regressing valence on time for each subreddit, 62% became more negative, while 38% became more positive **(Fig. 1c)**. Further separating subreddits by their year of creation shows that all subreddit cohorts showed a shift towards negativity (**Fig. 1d**). The cohort effect of negativity was independent from the within-cohort shift towards negativity. In other words, subreddits created later in time were more negative than those created earlier, and the same subreddits became more negative over time.

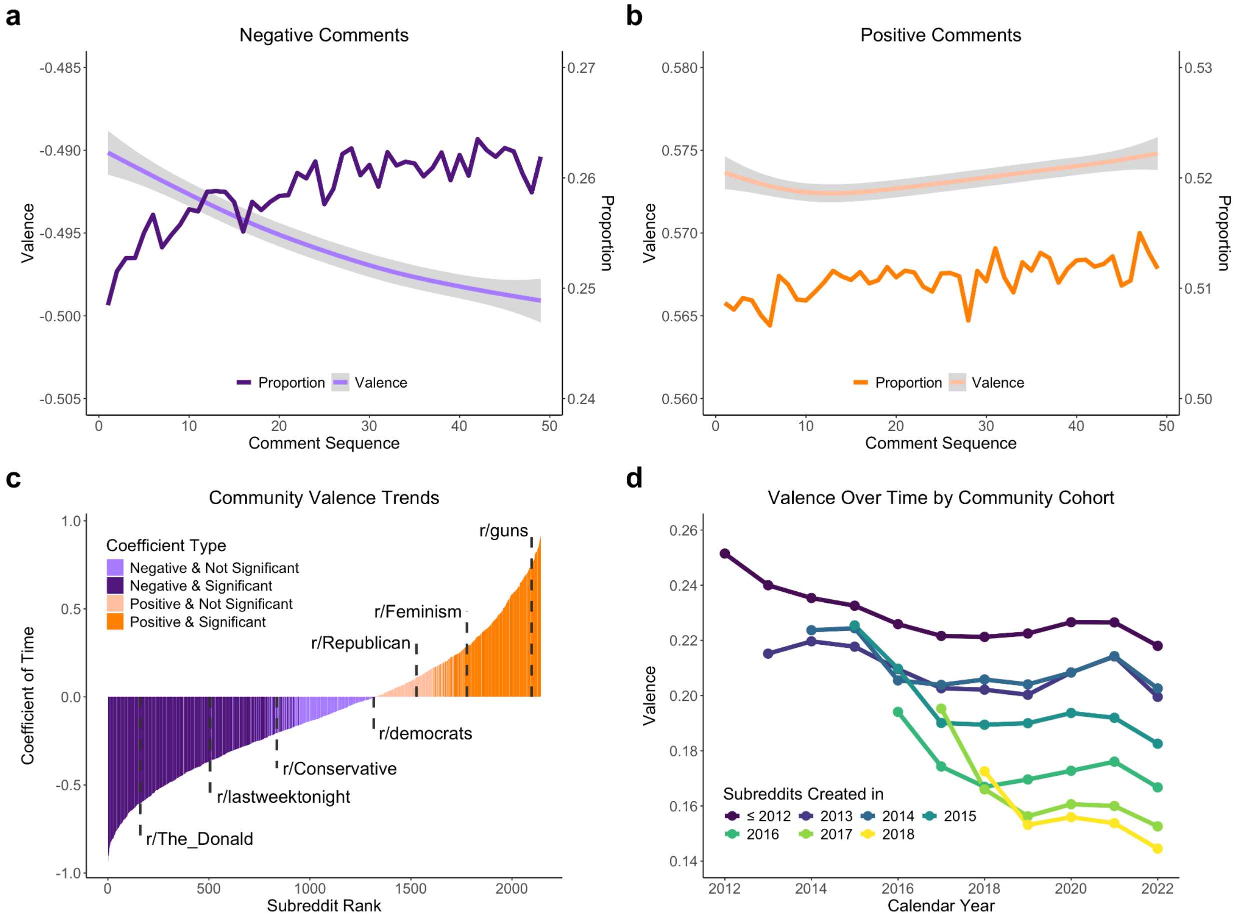

Fig. 1 | Valence change in communities and threads. a. Use of negative comments at different positions in threads. Negative comments were defined by the VADER classifier as those with negative valence scores. For a given sequence position, *proportion* refers to the percentage of negative comments among all comments at that position. *Valence* refers to the average valence score of negative comments at that position, estimated using a Generalized Additive Model (GAM). **b.** Use of positive comments at different positions in threads. Proportion and valence of positive comments were determined similarly to panel a. **c.** Each community's change in valence over time. For each subreddit, valence was regressed on time monthly, and the standardized coefficients were arranged in ascending order. Significance level is $p <$ 0.05. Eight subreddit communities were excluded for having fewer than 10 months of data. **d.** Valence change over time by community cohort. Data from 2012 to 2022 were used, and subreddit creation years were truncated at 2018 to visualize communities with at least five years of data. Subreddits created in or before 2012 were grouped together as one cohort. Monthly valence was aggregated by calendar year and cohort.

We also addressed alternative explanations for why valence may become more negative over time. One alternative explanation involves cohort replacement (i.e., new cohorts of users may be more negative than older cohorts). For example, there was an influx of politically polarized users around the 2016 election, which may have produced a shift towards negativity in 2016 (4). To investigate cohort replacement, we developed a method for decomposing new arrivals to a community each month and comparing their comment valence to that of existing users.

If our effect represents cohort replacement, then the negative trend should be driven by new users rather than old ones. However, we found that the decline in valence was stronger for old users, *b* = -0.04, *SE* = 0.003, *t* = -15.44, *p* < 0.001, *95% CIs* [-0.05, -0.04], than for new users, *b* = -0.02, *SE* = 0.003, *t* = -7.92, *p* < 0.001, *95% CIs* [-0.03, -0.02]. The negative trend in both old and new users suggests that both cohort replacement and endogenous change may be true. Alternatively, as suggested by previous research (27), it might be that new users become more negative because they quickly adapt to the changing norms among old users (see Supplementary Section 1.2.2 for more details).

A second alternative explanation is that our effect has nothing to do with social media. Instead, it may be that society has become more negative over the last decade. We address this alternative explanation by comparing our observed effects on Reddit with a time-matched digitized sample of 4.91 million news articles from 234 different publishers published between 2010 - 2022 from the News on the Web (NOW). We saw no trend towards negativity across news publishers, and we found that the monthly trend in valence among Reddit communities was significantly greater than the monthly trend in valence among news publishers (see Supplementary Section 1.2.3 for more details).

These findings were all consistent with the idea that negativity enables users to differentiate themselves as social media discussions become more saturated. To further explore this explanation, we analyzed the semantic differentiation of comments.

Semantic Differentiation as a Mechanism. Our second hypothesis was that the trend towards negativity should be explained by semantic differentiation. Specifically, comments should only become more negative over time to the extent that negative comments are semantically unique from what has already been said. We therefore tested whether semantic differentiation could account for the trend toward increasing negativity in online dialogue.

Within threads, we operationalized a comment's *semantic uniqueness* as its cosine distance from the thread's centroid embedding, calculated using a Sentence Transformer model. With this measure, we found that later comments in threads were more distinct than earlier comments in threads (**Fig. 2**), *b* = 0.06, *SE* = 0.0004, *t* = 160.56, *p* < 0.001, 95% CIs [0.062, 0.063]. Additionally, more negative comments were more semantically distinct, *b* = –0.04, *SE* = 0.0004, *t* = –88.04, *p* < 0.001, 95% CIs [–0.038, –0.036]. Crucially, when we controlled for semantic uniqueness, the previously observed trend toward greater negativity over time was no longer statistically significant, *b* = 0.0004, *SE* = 0.0004, *t* = –1.02, *p* = 0.308, 95% CIs [–0.0120, 0.0004]. This suggests that semantic differentiation underlies the observed increase in negativity over the course of online discussions.

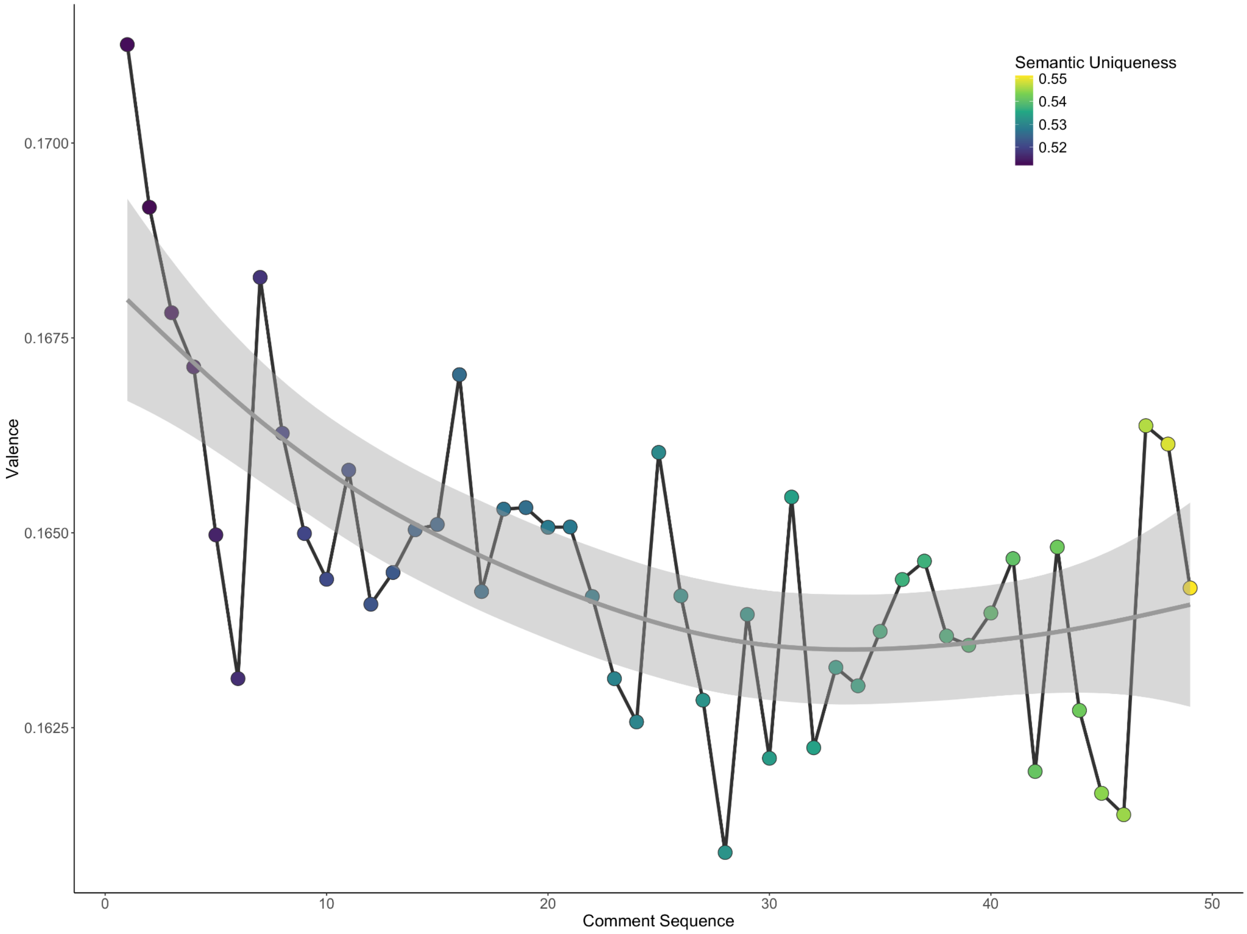

Fig. 2 | Valence at different positions in threads by semantic uniqueness. Data were aggregated by comment position, and valence trends were estimated using GAM.

We observed a similar pattern at the community level. At the community level, we translated semantic uniqueness to be semantic diversity—the degree of differentiation in comments within a community. We calculated the average semantic distance between each comment and the monthly centroid embedding, using the same SentenceTransformer model. The centroid represents the mean semantic content of all comments posted within a given month. We found that, as communities aged, they became more semantically diverse, *b* = 0.44, *SE* = 0.02, *t* = 25.89, *p* < 0.001, 95% CIs [0.40, 0.47]. Semantic diversity was also strongly associated with more negative community-level valence, *b* = –0.07, *SE* = 0.001, *t* = –39.20, *p* < 0.001, 95% CIs [–0.08, –0.07], and when we controlled for semantic diversity, the previously observed association between the passage of time and increasing negativity disappeared, *b* = 0.002, *SE* = 0.004, *t* = 0.55, *p* = 0.59, 95% CIs [–0.007, 0.01]. This supports the interpretation that the rise in negativity over time was statistically accounted for by the increasing semantic diversity of user contributions.

Both of these analyses supported the hypothesis that comments should only become more negative over time to the extent that negative comments are semantically unique from what has already been said. When controlling for semantic uniqueness, the effect of comment sequence on negativity was no longer significant, consistent with the hypothesis that semantic differentiation underlies the trend.

Moderation by Initial Valence. Our third hypothesis was that trends towards negativity should be moderated by the initial positivity of dialogue. Dialogue should trend more strongly towards negativity when it is very positively valenced, and less strongly when dialogue is less positive. When dialogue starts off positive, negative comments stand out in two ways—both because they are counter-normative and because there are more distinct ways to express negativity. In contrast, when dialogue is negative, both positive and negative comments stand out in different ways, making valence trends less pronounced. This may explain why we observed a rise in both negative and positive comment proportions over time in our initial analysis—both positive and negative comments stand out in different ways as dialogue becomes more negative.

To test this moderation hypothesis, we examined whether the strength of the negativity trend depended on a thread or community's initial valence. We saw evidence for this hypothesis in both threads and communities. Visualizations showed that subreddits with higher initial valence were more likely to exhibit a decreasing valence trend (**Fig. 3a**). When threads with negative or neutral initial valence showed an increase in valence across the sequence, those with positive initial valence showed the strongest decrease (**Fig. 3b**), mostly because of a rising proportion of negatively valenced comments and a declining proportion of positively valenced comments over time (**Fig. 3c**).

Using mixed-effect regressions, we found that the rise in negativity was steeper as a function of initial thread positivity, *b* = -0.04, *SE* = 0.0004, *t* = -102.03, *p* < 0.001, *95% CIs* [-0.042, -0.040], and initial community positivity, *b* = -0.04, *SE* = 0.001, *t* = -30.87, *p* < 0.001, *95% CIs* [-0.045, -0.043]. The decline in valence was significant and robust among threads, *b* = -0.04, *SE* = 0.0006, *t* = -77.16, *p* < 0.001, *95% CIs* [-0.042, -0.040], and communities, *b* = -0.07, *SE* = 0.003, *t* = -23.50, *p* < 0.001, *95% CIs* [-0.08, -0.07], with relatively high initial valence (1 SD above the mean). In contrast, valence increased among threads, *b* = 0.04, *SE* = 0.0006, *t* = 67.1, *p* < 0.001, *95% CIs* [0.037, 0.039], and communities, *b* = 0.01, *SE* = 0.003, *t* = 3.13, *p* = 0.002, *95% CIs* [0.004, 0.0160], with relatively low initial valence (1 SD below the mean).

In sum, threads and communities with more positive initial valence showed steeper declines in valence over time. This pattern was statistically robust and asymmetric: positivity tended to fall more sharply than negativity rose. This is consistent with our theory. Notably we also saw asymmetric valence convergence in our experiment (positive content grew steeply negative; negative content grew slightly positive).

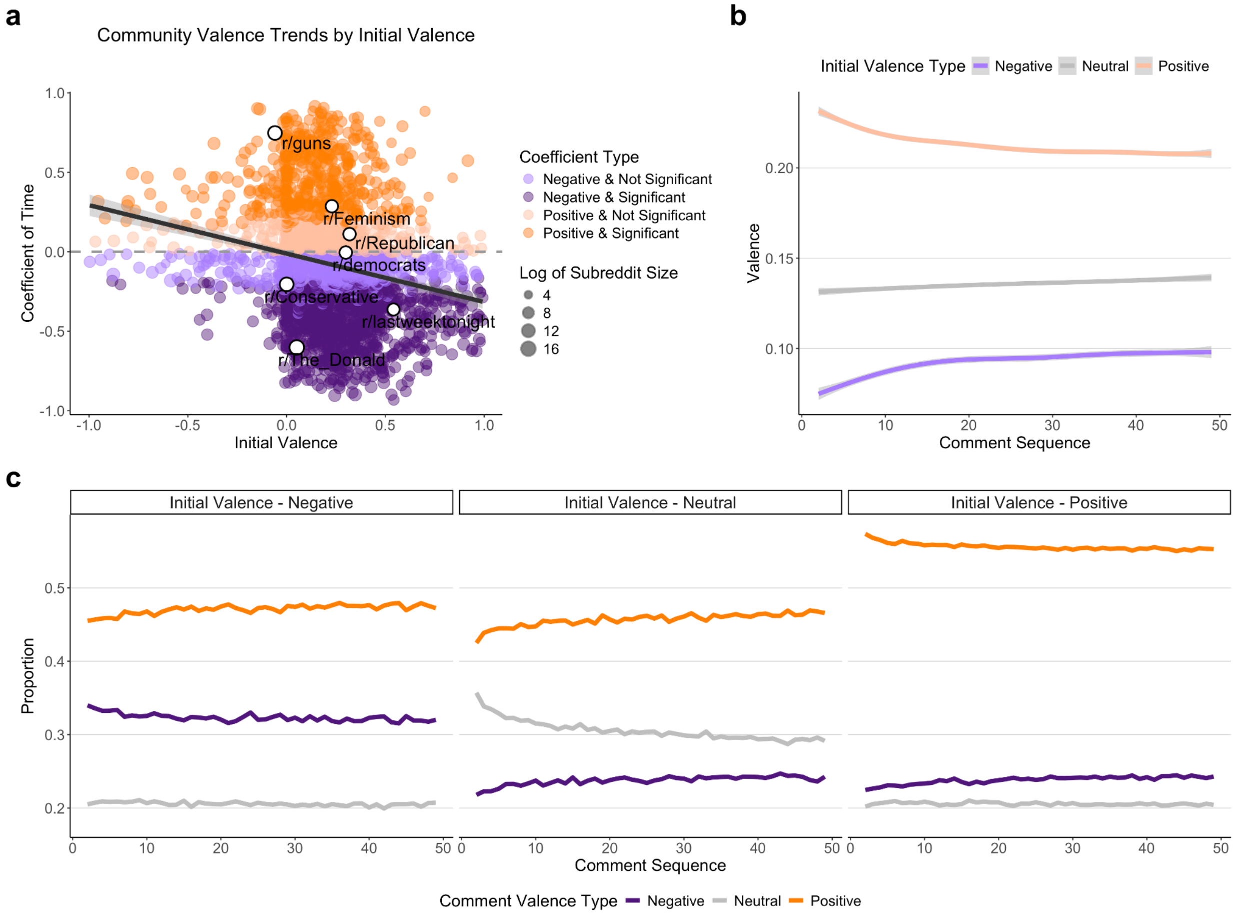

Fig. 3 | Valence change in communities and threads by initial valence. a. Each community's change in valence over time plotted against its initial valence. The initial valence of each subreddit community refers to the average valence of comments posted during its first month. Time coefficients and significance levels are the same as in **Fig. 1a.** Node sizes correspond to the log-transformed average monthly active users (base 2). **b.** Valence at different positions in threads by initial valence. A thread's initial valence was the valence of the first comment, which was classified by VADER into negative, neutral, or positive. Valence at each sequence position was estimated using GAM. **c.** Proportions of negative, neutral, and positive comments at different sequence positions, stratified by initial thread valence. Classification and estimation followed the same procedure as in panel **b**.

Another way to test the same hypothesis is to look at the functional form of declines in negativity, rather than testing for moderation by initial valence. If dialogue reaches a saturation point—where negativity becomes less novel and both negative and positive comments can serve as differentiating signals—we would expect the trend toward negativity to flatten or reverse. This would produce a quadratic trend over time. By adding a quadratic term for time to the previous model examining comment valence decline in communities, we found evidence that community valence tended to rebound towards positivity over time, *b* = 0.016, *SE* = 0.002, *t* = 8.44, *p* < 0.001, *95% CIs* [0.01, 0.02]. Similarly, including a quadratic term for sequence revealed that comments posted later in a thread may also show a rebound towards positivity, *b* = 0.004, *SE* = 0.0006, *t* = 6.47, *p* < 0.001, *95% CIs* [0.003, 0.005] (see **Fig. 2**). However, this effect was only observed for non–first-level comments (i.e., replies to other comments), not for first-level comments that directly reply to the original post, *b* = 0.001, *SE* = 0.0007, *t* = 1.57, *p* = 0.12, *95% CIs* [-0.0003, 0.0025]. Non–first-level comments may need to differentiate themselves from the comment they respond to, making rebounds in valence more likely.

Analyzing the Role of Community Size. Fourth and finally, we tested whether our previous effects would replicate with community size as the independent variable (defined here as the average number of monthly active users), rather than time. Our reasoning was that larger communities make it difficult for social media users to differentiate themselves for the same reason that time makes it difficult for social media users. In support of this logic, key effects replicated when we replaced time with community size.

As reflected in **Fig. 4a**, our model comparing communities suggested that comments in larger communities were more negatively valenced than comments in smaller ones, *b* = -0.16, *SE* = 0.02, *t* = -7.50, *p* < 0.001, *95% CIs* [-0.20, -0.12]. Another model focusing on changes within communities suggested that comments became more negative in months where there were more active users in a community, *b* = -0.013, *SE* = 0.001, *z* = -9.05, *p* < 0.001, *95% CIs* [-0.02, -0.01], even when controlling for the passage of time.

We also replicated the positive association in cross-correlations, which tested for temporal lags between variables while removing autocorrelation from the time series. In these models, increases in community size were most strongly associated with more negatively valenced comments within the same month, *b* = -0.05, *SE* = 0.002, *t* = -24.43, *p* < 0.001, *95% CIs* [-0.055, -0.047], compared to other months for the same subreddit (**Fig. 4b**). This same relationship replicated when we changed the time window to weeks rather than months (see Supplementary Section 1.2.6).

As with the effect of time, the effect of community size was explained by semantic differentiation, although this time the effect was only partially explained. Consistent with **Fig. 4a,** larger communities were more semantically diverse than smaller communities, *b* = 0.2, *SE* = 0.02, *t* = 8.35, *p* < 0.001, *95% CIs* [0.14, 0.22], and after controlling for semantic diversity, the effect of community size on negative valence was reduced by 28%, *b* = -0.11, *SE* = 0.02, *t* = -5.48, *p* < 0.001, 95% *CIs* [-0.15, -0.07]. Similarly, monthly increases in the number of active users were most strongly associated with more semantic diversity within the same month, *b* = 0.05, *SE* = 0.002, *t* = 26.38, *p* < 0.001, *95% CIs* [0.05, 0.06], compared to other months for the same subreddit (**Fig. 4c**). After controlling for semantic diversity, the effect of rising community size on negative valence was reduced by 19%, *b* = -0.011, *SE* = 0.001, *t* = -7.34, *p* < 0.001, 95% *CIs* [-0.013, -0.008].

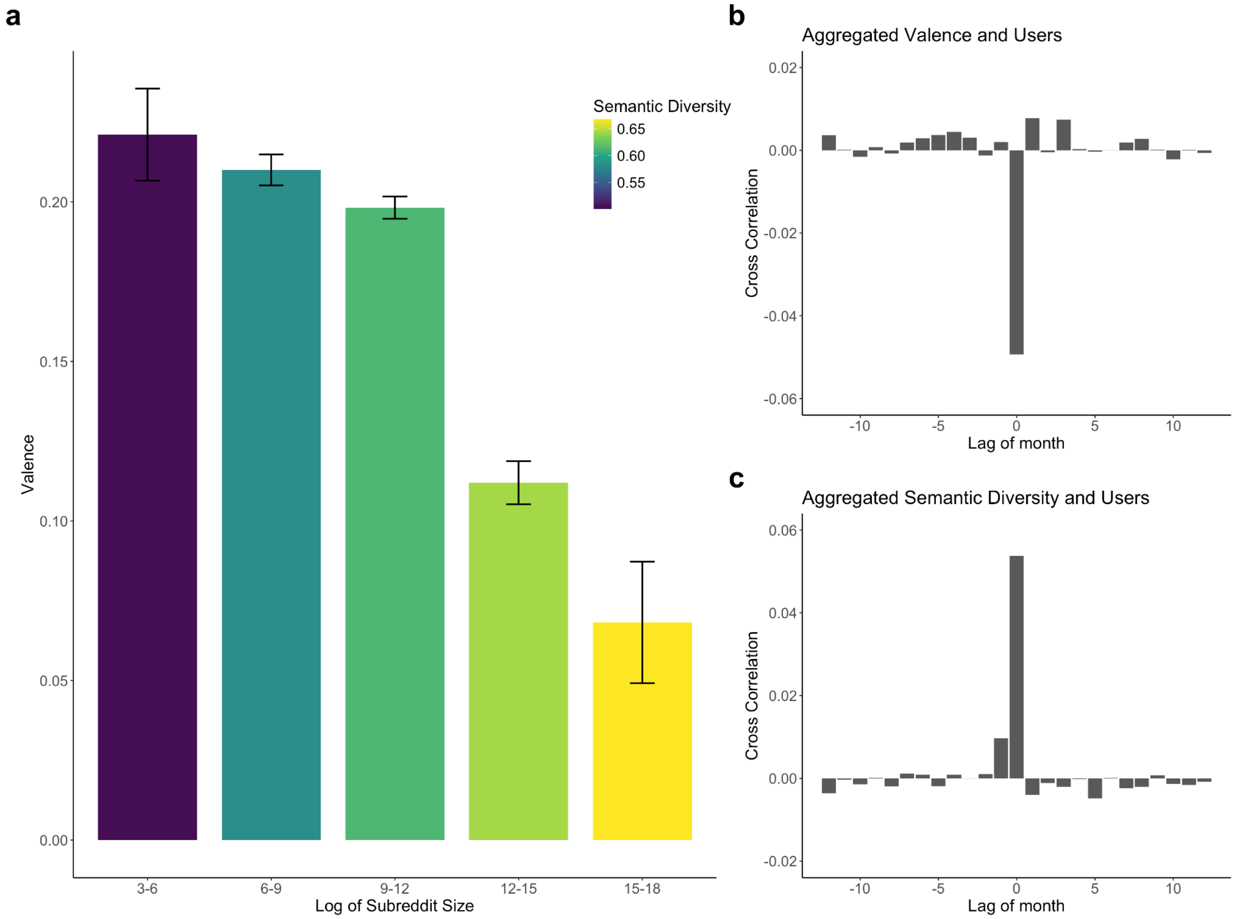

Fig. 4 | Associations between size, valence, and semantic diversity in communities. a. Subreddit valence by size and semantic diversity. Subreddits were grouped into five bins based on their average monthly active users, using log-transformed values (base 2). Each bin covers a $\log_2$ range of 3 units. For instance, *r/democrats*, with an average of 1,340 monthly users ($\log_2 \approx 10.4$), falls into the third bin: 9–12. **b.** Cross-correlation between valence and user count at different monthly lags. For example, a lag of -3 indicates correlation between the number of active users in the current month and valence three months later. Correlations were aggregated across subreddits. Error bars represent ±1 standard deviation. **c.** Cross-correlation between semantic diversity and user count at different monthly lags. Lag structure and aggregation followed the same procedure as in panel b.

Summary of Observational Findings. We found support for each of our four hypotheses. First, comments in threads and communities became more negative over time. Second, we found that this trend was statistically accounted for by semantic differentiation, defined as semantic uniqueness of comments in threads and semantic diversity of communities. Third, we found that this trend towards negativity was sublinear; it was largest when threads and communities started off with positively valenced comments, and it leveled off and eventually reversed among the most negative threads and communities. Fourth, we replicated key analyses using community size rather than the passage of time, based on the idea that community size makes differentiation more challenging for the same reason that the passage of time makes differentiation more challenging.

Although our observational analysis is large in scope and ecologically valid, it provides limited grounds for causal inference. We cannot definitively conclude that a motivation to differentiate caused users to post more negatively valenced or semantically unique comments. Other processes may unfold as threads grow longer or communities become larger. For instance, users may initiate communities and threads during moments of collective enthusiasm, and the observed decline in positivity over time may reflect a natural fading of that initial excitement. While our ability to replicate the effect using community size—a measure independent of time—makes this explanation less likely, we cannot fully rule it out. We are also limited by the fact that we cannot observe what users might have said before a thread or subreddit began. To address these limitations and test our theoretical mechanism more directly, we conducted a large, multi-generational experiment in which these confounds could be better controlled.

Experimental Evidence From a Multi-Generation Experiment

We designed a pre-registered experiment to test our first three hypotheses: (1) comments in an evolving thread would become more negatively valenced over time, (2) this trend would only manifest when users were incentivized to differentiate themselves, (3) this trend would be steepest when dialogue started out positive. One distinct advantage of running an experiment is that we could manipulate whether participants sought to differentiate themselves from other users through the incentive scheme.

Experiment Design. We recruited 4,000 participants from Prolific to participate in a five-generation study in which they would add comments to an evolving discussion thread of a news headline. Our goal was to create ecologically valid discussions that would systematically vary in their initial valence, but would be similar on other key dimensions. To do so, we curated four real news headlines by first scraping Google News and picking headlines that were published on the same day by outlets with similar impact to each other and low political relevance. We then had candidate article headlines rated by a large language model (GPT-4-turbo) on a scale from -1 (very negative) to 1 (very positive), and selected four headlines that were rated as positive ("A Japanese Organization of Atomic Bombing Survivors Wins the 2024 Nobel Peace Prize"), neutral ("Hundred-Year-Old Remains Of British Man Who Died On Everest Discovered By Nat Geo Doc Team Including, Free Solo's Jimmy Chin"), moderately negative ("Sealed TikTok court documents show time limit tool effectively did nothing to reduce teen usage, NPR reports")[1] and highly negative ("Hurricane Milton Live Updates: Death Toll Climbs as Florida Assesses Storm Damage").

We pre-registered using an LLM (GPT-4o) to rate valence because LLMs provide more contextually sensitive ratings than dictionary-based methods like VADER (28–30). We could not use an LLM in our Reddit analysis because the number of comments made it computationally prohibitive, but using them was feasible in our experiment.

Participants in each generation were randomly assigned to contribute to a pair of threads: either the threads about the positive and the neutral headline, or the threads about the moderately negative and highly negative headline. We created this assignment so that participants did not discuss extremely negative and extremely positive articles at the same time, since this could lead to affective spillover.

In each thread, participants scrolled through comments from the previous generation, and then contributed their own comment. Following past studies on differentiation motivation (13), we randomly assigned participants to either a differentiation condition or a conformity condition. In the differentiation condition, we incentivized participants to write comments that were unique from past comments in the thread. In the conformity condition, we incentivized participants to mimic past comments. Our incentive was a $5 reward would be given either for producing the most unique comment (in the differentiation condition) or for producing the most similar comment (in the conformity condition). Participants in both conditions had to use their own words to be eligible.

In sum, this experiment had a 5 (generation) x 2 (headline valence) x 2 (incentive type) design where all conditions were between-subjects. Translated to this design, our pre-

[1] The source, "NPR reports", was changed to "according to news reports" when displaying the news headline to participants.

registered hypotheses were that (1) participants' comments would become more negatively valenced over time, (2) this trend would only manifest when participants were incentivized to differentiate themselves, and (3) this trend would be most prominent in the positive headline valence condition.

Findings. To test our hypotheses, we first fit a regression in which LLM-rated comment valence was regressed on generation, headline valence condition, and incentivize type, and the interaction terms of these fixed effects. In this regression, we found evidence for our hypothesized three-way interaction, $b = 0.06$, $SE = 0.01$, $t = 4.68$, $p < 0.001$, *95% CIs* [0.04, 0.09]. When participants were incentivized to differentiate themselves, generation interacted with headline valence, $b = -0.03$, $SE = 0.01$, $t = -3.69$, $p < 0.001$, *95% CIs* [-0.05, -0.02]. Specifically, posts about the most positive headline had the most salient downward valence trend over generations, $b = -0.10$, $SE = 0.03$, $t = -3.49$, $p <$ 0.001, *95% CIs* [-0.16, -0.04]. The effect was smaller but still statistically significant for the neutral headline, $b = -0.05$, $SE = 0.02$, $t = -2.25$, $p = 0.02$, *95% CIs* [-0.10, -0.01], was marginally significant for the moderately negative headline, $b = -0.03$, $SE = 0.01$, $t =$ -1.89, $p = 0.06$, *95% CIs* [-0.06, 0.001], and did not reach statistical significance for the highly negative headline, $b = 0.004$, $SE = 0.02$, $t = 0.21$, $p = 0.83$, *95% CIs* [-0.03, 0.04].

For participants who were incentivized to conform, we found a very different pattern of results, in which valence and generation interacted positively rather than negatively, $b =$ 0.03, $SE = 0.01$, $t = 2.92$, $p = 0.003$, *95% CIs* [0.01, 0.04]. In the incentivized conformity condition, we found that the posts about the positive and neutral headlines had upward valence trends over generations: the neutral headline showed a strong effect, $b = 0.14$, $SE = 0.02$, $t = 6.45$, $p < 0.001$, *95% CIs* [0.10, 0.19], and the positive one showed a marginally significant effect, $b = 0.06$, $SE = 0.02$, $t = 2.55$, $p = 0.01$, *95% CIs* [0.01, 0.10]. In contrast, posts about the two negative headlines had no valence trends over generations: the moderately negative headlines did not show a statistically significant effect, $b = -0.02$, $SE = 0.01$, $t = -1.59$, $p = 0.11$, *95% CIs* [-0.04, 0.004], and the highly negative headline showed an even weaker effect, $b = 0.01$, $SE = 0.01$, $t = 0.91$, $p =$ 0.36, *95% CIs* [-0.01, 0.04]. All of these effects and the detailed distributions of response valence ratings are illustrated in **Fig. 5a-d**.

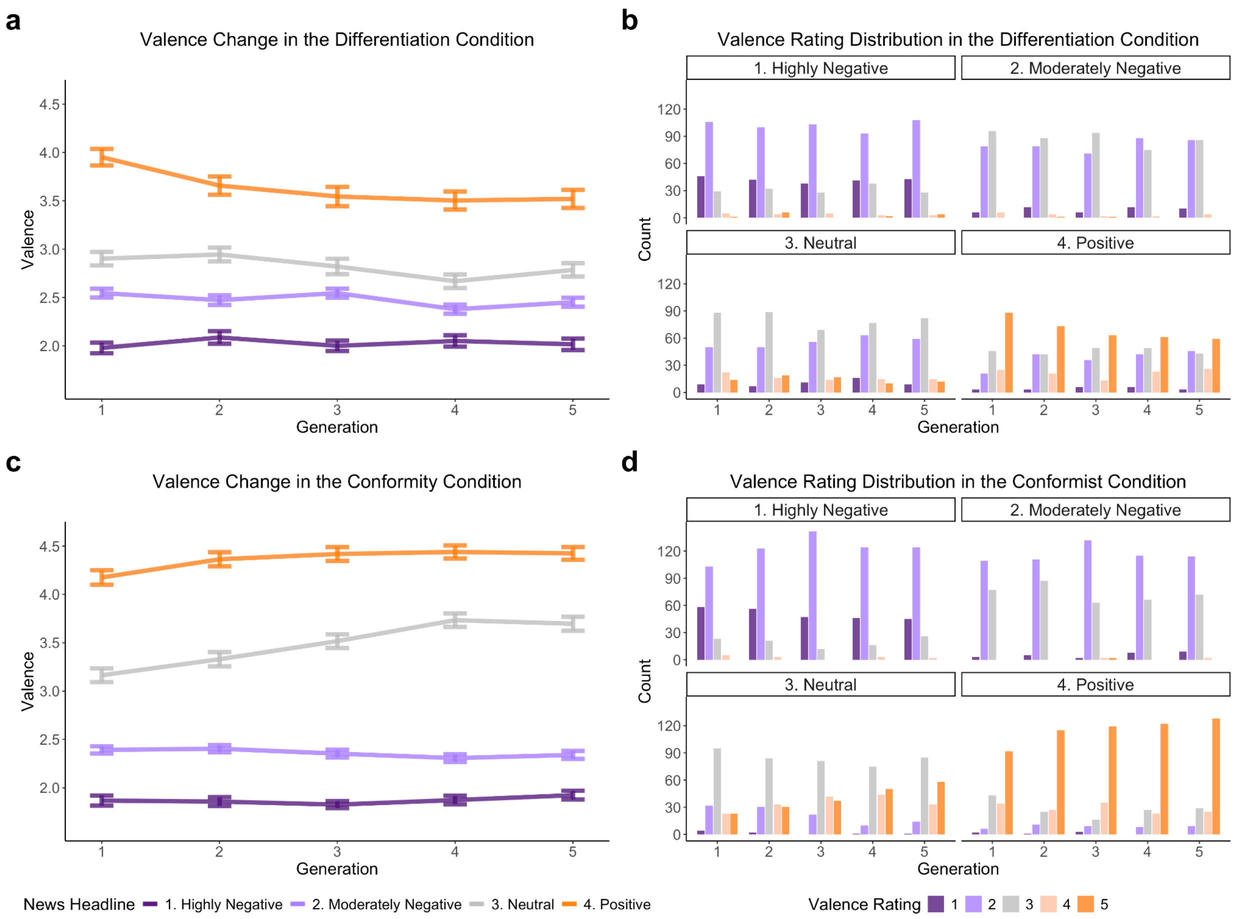

Fig. 5 | Valence change across generations in conformity and differentiation conditions. a. Valence of participants' responses to different news headlines across generations in the differentiation condition. Headlines correspond to varying levels of initial valence. Error bars represent ±1 standard deviation. **b.** Distribution of valence ratings across generations in the differentiation condition. **c.** Valence of participants' responses to different news headlines across generations in the conformity condition. Headlines and error bars are the same as in panel a. **d.** Distribution of valence ratings across generations in the conformity condition.

In sum, we found robust support for each of our three hypotheses. Moreover, we observed similar effects in a supplementary experiment (*n* = 1000) in which we gave participants no explicit incentive to differentiate themselves or to conform (see Supplementary Section 2.3). The supplementary experiment suggests that people may spontaneously try to differentiate themselves when they contribute to discussions on social media. With random assignment of participants to valence, time, and incentive, our experiments offer causal evidence that incentives to differentiate contribute to valence decline over time, consistent with our broader motivated differentiation account.

Discussion

Across 2,150 Reddit communities including 2.05 billion comments, and in a large multi-generational experiment, comments on social media became more negative over time. Three patterns in our data suggest that this trend arose at least partly as a byproduct of users' motivation to differentiate themselves.

First, the trend towards negativity was explained by semantic differentiation. Using a well-established deep learning method of capturing semantic similarities and differentiation (31), we found that negative comments had higher levels of semantic differentiation, and that controlling for semantic differentiation statistically either partially or fully explained the trend in negativity over time, depending on the exact analysis. In our experimental dataset, we manipulated participants' motivation to either differentiate themselves or to conform to norms through monetary incentives. We only found a trend towards negativity when participants were incentivized to differentiate themselves.

Second, the trend towards negativity was moderated by the positivity of dialogue. It was steepest when conversations were highly positive, but it flattened or reversed when conversations became more negative. This flattening was not because people became less likely to leave negative comments (a classic "floor effect" explanation). Rather, it was because the rise in negative comments was matched with a rise of positive comments. We suggest that, as discourse becomes more negative, both negative and positive comments allow users to differentiate themselves: positive comments are unique in negative threads because they are counter-normative, whereas negative comments can be unique because negative information is heterogeneous; there are many unique ways to express negativity.

Third, we found that group size showed the same effects of time in our observational dataset. We view group size as theoretically similar to time. Just as users later in a discussion have more pressure to differentiate themselves than users earlier in a discussion, users in larger communities have more pressure to differentiate themselves than users in smaller communities. Consistent with this prediction, larger communities were more negative than smaller communities, changes in community size were associated with shifts towards negative dialogue, and both of these effects were statistically explained by semantic uniqueness. Our analyses disentangle any covariation between community size and age to ensure that these are independent predictors of negativity, rather than confounded predictors.

A key question is whether users are actually rewarded for posting unique comments. It is challenging to explore this question using our observational data because Reddit only provides information about each comment's total score (the sum of upvotes and downvotes). In Supplemental Section 1.2.8, we show that semantic differentiation has a quadratic relationship with score: Moderately unique comments receive higher scores than comments that are not at all or extremely unique. However, this test is flawed because highly unique comments could receive engagement in many other ways (e.g., eliciting a high volume of upvotes as well as downvotes, or receiving a large share of replies). Despite several past studies showing that social media users want to be unique (16,17), there are curiously few thorough tests of whether unique content provides users with the rewards that they expect. We view this as an important area for future research.

Our supplementary materials also include a range of other analyses, such as testing whether these effects characterize both political and apolitical social media communities, and testing whether the shift towards negativity also characterizes online news sites. We show that our effects generalize across both kinds of social media community, but do not generalize to the news. This is consistent with our theory, since the news does not meet our key theoretical criteria (people making sequential comments on the same topic in an evolving discussion thread). The fact that our effect generalizes across political and apolitical communities is important, because it suggests that motivated differentiation is a fundamental mechanism that can make social media dialogue more negative even when intergroup competition is not salient.

One interesting feature of our observational results was that dialogue never became highly negative at the aggregate level, even after valence declined over time. Rather than crossing the neutral “0” point of valence, we found an equilibrium around 0.20. In our view, this equilibrium reflects the fact that human perceptions of neutrality are seldom aligned with true neutrality. Negative information is more dominant than positive information, which means that dialogue must be more extremely positive to be subjectively experienced as positive, whereas dialogue can be mildly negative to be subjectively experienced as negative (26,32). Comments like “Sure, I guess you’re right” and “I’m fine” may likewise receive mildly positive scores in a sentiment classifier such as VADER, but would be perceived as negative by users. Readers should therefore treat a neutral “0” point in a valence classifier with caution. In our experimental studies, we used a more contextually sensitive LLM to classify valence, which revealed more outwardly negative dialogue.

The most troubling aspect of our findings is that they suggest that negativity on social media could be relatively robust to intervention. Intergroup animosity may improve when algorithms explicitly promote content that “bridges” the preferences of liberal and conservative users (33,34), and algorithms that promote prosocial content may alleviate users’ misperceptions that moral outrage is normative in online social networks (35). However, it is more difficult to disincentive the motive to be unique on social media. Furthermore, our research suggests that algorithms that promote positive and prosocial content may even have unintended backlashes because it incentivizes negative dialogue by making it even more unique. Differentiation-based negativity may prove to be a fundamental obstacle to sustaining positive dialogue on social media.

Our studies had limitations that are important to acknowledge. First, we only analyzed content on Reddit, which has features that are not shared by other online social networks. For example, most communities on Reddit are politically liberal, although we show that political and apolitical communities showed similar effects in our supplemental materials (Supplementary Section 1.2.7). More importantly, Reddit is anonymous, so users may feel more comfortable sharing negative comments because they feel less pressure to agree than on more intimate platforms like Facebook or Instagram. Users in our experiments were also anonymous. A key goal for future studies will be to replicate our analyses across a wide variety of platforms that vary in their structural qualities (e.g., anonymity, community organization, demographic characteristics, thread organization) to test the generalizability of differentiation-based negativity. One recent paper analyzed toxicity on many different social media platforms over time, but presented descriptive results rather than testing a psychological theory of why dialogue might change over time (6).

Second, there were key differences between our experiments and our observational data analysis: while users on Reddit can return to threads and communities to continue dialogue at any point, our experimental participants only contributed comments at one time-point, and could never engage with one another in a sustained discussion. Furthermore, Reddit includes different tiers of response (e.g., level-1 comments focus on the original post; higher-level comments focus on previous comments), whereas comments in our experiments only focused on the post. These limitations were somewhat inevitable to ensure that users commented in the same format across generations, but they represent key differences between our paradigms.

These limitations suggest that the narrowest version of our account is that differentiation-based negativity is likely to plague online social networks with anonymous users and topic-based communities and threads, whereas the broadest version of our account is that differentiation-based negativity plagues all online social networks. Either way, our research suggests that differentiation-based negativity affects the experience of millions of social media users, and represents a fundamental and persistent barrier to sustained positive engagement. As long as social media websites represent attention economies, they may also incentivize disagreement and division.

Methods

Observational Analyses

Reddit Data Collection. We extracted data from a Reddit dump collected by developers via Pushshift[2] (36). Currently, the dump has all Reddit submissions and comments from June 2005, when Reddit was established, to December 2024. By the time the dump was downloaded for analysis, it had been updated through February 2023. We focused particularly on a sample of 2,150 subreddits which had been annotated by prior research in terms of their volume of political discussion. A subset of 1,149 of our subreddits had also been annotated by past studies based on their partisanship (4). We removed non-English subreddits from our sample, since our measures were based on English-language words. Supplementary Section 1.1.1 has more information about how we sampled subreddits.

After excluding removed or deleted comments (which cannot be identified in the dataset), our dataset contained 2.05 billion comments, accounting for approximately 15% of all activity on Reddit. The oldest subreddits in our sample, namely r/obama and r/entertainment, were established in 2007, while the vast majority were established on or after 2008. For robust analysis of changes within and variations between subreddits, we only considered subreddits established in and after 2008 in the study. The full list of subreddits used is available at https://osf.io/68wvm/.

We further sampled 115,936 threads from the selected subreddits for our thread-level analysis. To ensure that we had enough comments to examine in-thread valence change, we limited our selection to threads with more than 50 comments, yielding 115,979 threads. We collected all identifiable comments belonging to these threads, resulting in approximately 10.5 million comments. Some threads included comments without valid information. Removing these threads yielded 115,936 threads. Notably, thread lengths followed an exponentially decreasing distribution, with only a handful of threads being extremely long. For consistency, we truncated each thread to include only the first 50 comments. Additionally, because the first comment in many threads was an auto-generated bot message, we removed the initial comment from every thread. The final dataset contains approximately 5.5 million comments.

Semantic Diversity and Uniqueness. Semantic diversity represents the breadth of different topics within a subreddit. Following previous research (37), we calculate semantic diversity within a particular time-span (a monthly basis). The semantic diversity of a comment is calculated as the average cosine distance between the comments embedding and the centroid of comment embeddings for that month in the embedding space. If comments are widely dispersed around the centroid, the average

[2] Users who submitted the form to request for their accounts being removed from the PushShift API were not included in the dump and thereby not in our study.

cosine distance will be high, indicating high semantic diversity. Comments' high-dimensional embeddings were obtained using a sentence transformers model[3].

We calculated each comment's semantic uniqueness within threads by embedding every comment using a sentence transformer and obtaining the centroid of all comments. Next, we calculated the cosine distance of a comment's centroid to the centroid of the thread as a measure of semantic uniqueness.

Valence. We used VADER, a dictionary-based text classifier, to measure comment valence (38). VADER yields a compound score ranging from -1 to 1 for comments, where 1 means very positive and -1 means very negative. We take the average of VADER valences of all comments within a month as the representation of valence of the subreddit for the month. In our supplementary materials, we measured valence in the NOW dataset using a similar approach. We first measured the valence of news headlines, then aggregated the results by publisher-month pairs.

Subreddit Age. A subreddit's age, or the date of creation, was determined based on the timestamp of its first appeared comment in our dataset. We used the year and month information in the timestamp as the approximation of subreddits' establishment time. Using this approach, the subreddits used in the study were identified as being established between 2008 and 2019.

Subreddit Size. We measured the monthly number of users in each subreddit by counting the unique authors of valid comments posted during that month. Then, subreddit size was calculated as the average number of monthly active users.

Comment Sequence. Within threads, we track timestamps of every comment and label them in a chronological order starting 0 to measure sequence. The first comments in threads were removed due to a high proportion of auto-generated bot comments (e.g., reminders of community rules).

Comment Level. Within threads, comments that directly reply to the submission post are marked as first-level comments. Comments that reply to other comments are considered higher-level comments.

Analysis. All of the regressions in our main text were mixed effects models in which we modeled intercepts and slopes as varying across subreddits. The thread analyses were 3-level models in which intercepts and slopes were modeled as randomly varying across threads which were in turn nested in subreddits.

To avoid cohort effects in our longitudinal analyses, we decomposed within- from between-subreddit effects using centering. In the main text, we report within effects because our hypotheses focus on how negativity has risen over time within

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

communities and threads, rather than cohort change. Our supplementary tables also contain the between effects.

Experimental Analyses

Pre-Registration. We pre-registered the sample size, measures, and analyses for this study. Our pre-registration is available at https://aspredicted.org/gmnm-xzkk.pdf.

Participants. We recruited 4000 participants, which was informed by a power analysis described in our supplementary materials. Our supplemental materials include complete information about the demographic characteristics of participants. All participants were recruited on Prolific, and paid $0.75 for their participation. We screened participants so that they could not participate in multiple generations of the study.

Stimuli. First, participants in our experiments were exposed to 20 randomly selected responses from participants in the previous generation under the same conditions while reading the news headlines. Second, the experiment included another stimulus to prompt either a conformist or differentiation motive. For the differentiation condition, participants were prompted to be as unique as possible from the perspectives of what other people have posted in their own words. For the conformist condition, participants were prompted to copy the perspectives of what other people have posted as closely as possible in their own words. Finally, We instructed participants in both conditions that we had developed a computer algorithm that could detect the similarity of social media posts. In turn, participants in the differentiation incentive condition read that we would give a $5 reward to the participant who wrote the post that was most *different* to what others said while staying on the topic of the headline, whereas participants in the conformity incentive condition read that we would give a $5 reward to the participant who wrote the most *similar* to what others said while staying on the topic of the headline. These instructions were honest. We used GPT4 to rate headline similarity and gave $5 rewards to the winning participants of each condition and generation. Our supplementary materials provide the complete setup for the experiment.

Procedure. Since our experiment used a generational design, we implemented a process to collect responses from a previous generation and feed them into the next generation. In the experiment, participants viewed either two positive or two negative news headlines, along with responses from participants in a previous generation who had been assigned to the same motivational condition. Participants in both the conformist and differentiation conditions in the first generation saw responses randomly drawn from the same seed survey with no motivational prompts.

After collecting responses, we used a GPT-4 API to (1) verify logical coherence, (2) check topic relevance, and (3) correct typos in coherent and relevant responses. Responses that were logically incoherent or irrelevant to the topic were removed.

Cleaned responses then formed the pool from which responses were drawn for the next generation. After completing five rounds of collection, we combined all responses to create the final dataset. Our supplementary materials provide the GPT4 prompts used for cleaning, filtering, and valence evaluation.

Measurement of Valence. Participants' responses were evaluated by GPT4 on a scale from 1 to 5 (1 = very negative, 3 = neutral, 5 = very positive). We opted out of using VADER to measure valence in participants' responses. VADER is a good tool for large scale text processing tasks for its decent performance and high efficiency. However, VADER as a dictionary-based method is not context-aware (29,30). In the context of experiments, where only a small number of responses toward a few pre-selected news were collected, context-specific news topics and some distracting words in responses could make VADER ratings potentially biased and less robust. Therefore, we used GPT4 to encode positivity to overcome those biases. We provide the complete prompt in our supplementary materials. We also show in Supplementary Section 2.2.3 that our experimental results were largely replicated using VADER.

Analysis. We analyzed our data using mixed effects models, modeling intercepts as varying randomly across participant ID.

Code and Data Availability

Analysis scripts and data are available at https://osf.io/68wvm/. Data and code may not be used for commercial purposes.

References

1. Fariello G, Jemielniak D, Sulkowski A. Does Godwin’s law (rule of Nazi analogies) apply in observable reality? An empirical study of selected words in 199 million Reddit posts. New Media Soc. 2024 Jan;26(1):389–404.

2. Van Bavel JJ, Robertson CE, del Rosario K, Rasmussen J, Rathje S. Social Media and Morality. Vol. 75, Annual Review of Psychology. Annual Reviews; 2024. p. 311–40.

3. Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ. Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci. 2017;114(28):7313–8.

4. Waller I, Anderson A. Quantifying social organization and political polarization in online platforms. Nature. 2021;600(7888):264–8.

5. Brady WJ, Crockett MJ, Van Bavel JJ. The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. Perspect Psychol Sci. 2020;15(4):978–1010.

6. Avalle M, Di Marco N, Etta G, Sangiorgio E, Alipour S, Bonetti A, et al. Persistent interaction patterns across social media platforms and over time. Nature. 2024;628(8008):582–9.

7. Rathje S, Robertson C, Brady WJ, Van Bavel JJ. People think that social media platforms do (but should not) amplify divisive content. Perspect Psychol Sci. 2024;19(5):781–95.

8. Auxier B. 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today [Internet]. Pew Research Center. 2020 [cited 2025 Apr 19]. Available from: https://www.pewresearch.org/short-reads/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/

9. Tokita CK, Guess AM, Tarnita CE. Polarized information ecosystems can reorganize social networks via information cascades. Proc Natl Acad Sci. 2021;118(50):e2102147118.

10. McLoughlin KL, Brady WJ, Goolsbee A, Kaiser B, Klonick K, Crockett MJ. Misinformation exploits outrage to spread online. Science. 2024;386(6725):991–6.

11. Brady WJ, Jackson JC, Lindström B, Crockett MJ. Algorithm-mediated social learning in online social networks. Trends Cogn Sci. 2023;27(10):947–60.

12. Leonardelli GJ, Pickett CL, Brewer MB. Chapter 2 - Optimal Distinctiveness Theory: A Framework for Social Identity, Social Cognition, and Intergroup Relations. In: Zanna MP, Olson JM, editors. Advances in Experimental Social Psychology [Internet]. Academic Press; 2010 [cited 2025 June 1]. p. 63–113. Available from:

https://www.sciencedirect.com/science/article/pii/S0065260110430026

13. Imhoff R, Erb HP. What Motivates Nonconformity? Uniqueness Seeking Blocks Majority Influence. Pers Soc Psychol Bull. 2009 Mar;35(3):309–20.

14. Berger J, Milkman KL. What Makes Online Content Viral? J Mark Res. 2012 Apr 1;49(2):192–205.

15. Kim HS. Attracting views and going viral: How message features and news-sharing channels affect health news diffusion. J Commun. 2015;65(3):512–34.

16. Drążkowski D, Pietrzak S, Mądry L. Temporary change in personality states among social media users: Effects of Instagram use on Big Five personality states and consumers' need for uniqueness. Curr Issues Personal Psychol. 2022;10(1):32–8.

17. Zeng X, Wei L. Social Ties and User Content Generation: Evidence from Flickr. Inf Syst Res. 2013 Mar;24(1):71–87.

18. Puryear C, Brady W, Jackson J, Leong Y, Kteily N. Rising Moralization in Social Media Discourse [Internet]. OSF; 2025 [cited 2025 Sept 21]. Available from: https://osf.io/z7p3u_v1

19. Alves H, Koch A, Unkelbach C. Why good is more alike than bad: Processing implications. Trends Cogn Sci. 2017;21(2):69–79.

20. Koch A, Alves H, Krüger T, Unkelbach C. A general valence asymmetry in similarity: Good is more alike than bad. J Exp Psychol Learn Mem Cogn. 2016;42(8):1171.

21. Averill JR. On the Paucity of Positive Emotions. In: Blankstein KR, Pliner P, Polivy J, editors. Assessment and Modification of Emotional Behavior [Internet]. Boston, MA: Springer US; 1980 [cited 2025 June 30]. p. 7–45. Available from: http://link.springer.com/10.1007/978-1-4684-3782-9_2

22. Jackson JC, Lindquist K, Drabble R, Atkinson Q, Watts J. Valence-dependent mutation in lexical evolution. Nat Hum Behav. 2023;7(2):190–9.

23. Alves H, Koch A, Unkelbach C. My friends are all alike—the relation between liking and perceived similarity in person perception. J Exp Soc Psychol. 2016;62:103–17.

24. Iliev R, M Bennis W. The Convergence of Positivity: Are Happy People All Alike? J Happiness Stud. 2023;24(5):1643–62.

25. Denrell J. Why most people disapprove of me: experience sampling in impression formation. Psychol Rev. 2005;112(4):951.

26. Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD. Bad is stronger than good. Rev Gen Psychol. 2001;5(4):323–70.

27. Rajadesingan A, Resnick P, Budak C. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In: Proceedings of the international AAAI conference on web and social media [Internet]. 2020 [cited 2025 July 1]. p. 557–68. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/7323

28. Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci. 2024;121(34):e2308950121.

29. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J Lang Soc Psychol. 2010 Mar;29(1):24–54.

30. Wilkerson J, Casas A. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annu Rev Polit Sci. 2017 May 11;20(1):529–44.

31. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Internet]. arXiv; 2019 [cited 2025 Sept 8]. Available from: http://arxiv.org/abs/1908.10084

32. Gottman JM, Levenson RW. The Timing of Divorce: Predicting When a Couple Will Divorce Over a 14-Year Period. J Marriage Fam. 2000 Aug;62(3):737–45.

33. Levy R. Social media, news consumption, and polarization: Evidence from a field experiment. Am Econ Rev. 2021;111(3):831–70.

34. Törnberg P, Valeeva D, Uitermark J, Bail C. Simulating social media using large language models to evaluate alternative news feed algorithms. ArXiv Prepr ArXiv231005984. 2023;

35. Brady WJ, Elnakouri A, Fatmi A, Finkel E, Jackson JC, Kteily N, et al. (Accepted in principle). Does algorithmic amplification cause people to misperceive norms during a national election? Nature.

36. stuck_in_the_matrix, Watchful1, RaiderBDev. Reddit Comments/Submissions 2005-06 to 2024-12 [Internet]. Available from: https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13

37. Shi F, Teplitskiy M, Duede E, Evans JA. The wisdom of polarized crowds. Nat Hum Behav. 2019;3(4):329–36.

38. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media. 2014. p. 216–25.

Supplementary Materials

1. Observational Analyses

1.1. Data Preprocessing

1.1.1. Sampling Subreddits

The sample of subreddits consists of two parts. The first part includes 1,149 subreddits sampled from a list of 10,086 subreddits with partisan scores (1). Subreddits' partisan scores were calculated based on membership similarity to seed partisan subreddit pairs such as r/democrats and r/Conservative. We split the partisan score spectrum from left to right into 10 quantiles and applied the following strategy to sample approximately 10% of the data while ensuring broad partisan score coverage: for every quantile, we kept all subreddits if the number within the quantile was fewer than 200, and randomly sampled 200 if more than 200 subreddits fell within the range. The 1,149 subreddits were established between 2007 and 2017. However, only r/obama and r/entertainment were established in 2007, and therefore we removed them from our sample. The rest of the subreddits sampled from this source were established in 2017 or earlier, which might not accurately reflect potential changes driven by events such as the 2016 presidential election.

Therefore, we sampled another 1,070 subreddits from a more comprehensive list of 30,899 subreddits (2), which included subreddits created between 2008 and 2019. While subreddits from this list do not have partisan scores, they do have an estimated proportion of politically related comments relative to all comments posted in the community. After excluding the overlap between subreddits with both an estimated proportion of political comments and those with partisan scores, we ranked the remaining 20,893 subreddits (those with only an estimated proportion of political comments) by size and selected the top 5%. We confirmed this cutoff by comparing the average size of subreddits above the cutoff with the average size of subreddits with partisan scores, yielding 1,070 complementary subreddits. These 1,070 subreddits account for approximately 40% of the activity of the remaining 20,893 subreddits.

Altogether, this sample contains 2,219 subreddits. However, it includes subreddits where the main communication language is not English, which are not compatible with our analytical plan. In the next section, we illustrate how we removed non-English subreddits to obtain the final sample.

1.1.2. Removing Non-English Subreddits

We applied the following rules to identify and remove non-English subreddits:

- First, if a token is neither in the English words set from the Natural Language Toolkit (NLTK) nor numeric, it is considered a non-English token.
- Second, if a comment contains more than 50% non-English tokens, it is considered a non-English comment.
- Finally, if a subreddit has more than 50% non-English comments, it is considered a non-English subreddit.

Following this procedure, we removed 67 non-English subreddits, leaving 2,150 effective subreddits for analysis. The list of removed subreddits is provided below.

1.1.3. Removed Subreddits

The removed subreddits were: wc2010_crests, latvia, UserSimulator, PERU, lithuania, the_schulz, AnimalsWithoutNecks, de_IAmA, FuckMarryOrKill, swedishproblems, bulgaria, Republica_Argentina, Fahrrad, rance, SpainPolitics, Belgium2, TurkeyJerky, GermansGoneWild, chonglangTV, nhaa, merval, Finanzen, farialimabets, newsokuexp, espanol, Ni_Bondha, AskFrance, China_irl, SubredditSummaryBot, Monterrey, futebol, BrasildoB, vosfinances, newsNepal, liberta, RepublicaArgentina, brasilnoticias, BrasilSimulator, EcuadorNoticias, dankgentina, wasletztepreis, brasilivre, ani_bm, Panama, Pikabu, RLCustomDesigns, kfq, SpotifyPlaylists, de_EDV, MAAU, Italia, burdurland, ItaliaPersonalFinance, investimentos, FreeDutch, Mujico, AthleticBabes, Suicidal_Insanity, tokkiefeesboek, Onlyfans_Promo, einfach_posten, fcporto, PORTUGALCARALHO, Thread_crawler, desabafos, KGBTR, OnlyFans101

Most of the removed subreddits communicate in languages other than English, such as Spanish, Portuguese, French, or Turkish. Note that some subreddits were also removed even though they appear to communicate in English. This is because they contain few informative posts; for example, they may consist largely of promotion codes or contain very few comments beyond emojis.

1.1.4. Sampling Threads

We selected the 115,961 threads used for finer-grained analysis with the following procedure: First, we excluded inactive subreddits from our sample. An inactive subreddit is operationalized as one with fewer than 12 months containing more than 30 submissions per month. Submissions are defined as the original posts that start a thread and receive replies from other users. Second, for each subreddit in our final sample, we collected all submissions with a sufficient and reasonable number of replies.

We applied an adapted interquartile range (IQR) method to determine which submissions to keep in a subreddit: (1) submissions had to have at least 50 replies to ensure sufficient length for sequence analysis, and (2) submissions' replices could not exceed the upper bound—defined as the 75th percentile plus 1.5 times the IQR—or 10,000 replies if no threads exceeded the upper bound. The additional 10,000 hard limit was set to remove unusually large threads in small communities, which are untypical. Using this approach, we obtained 115,979 threads from 2,046 subreddits. Finally, we removed comments without identifiable positions in threads (i.e., comments where the replied-to post could not be determined), leaving a final set of 115,961 threads.

1.1.5. Measuring Valence

We used VADER to obtain the valence of every comment. VADER is a reliable and efficient rule-based model suitable for measuring sentiment in microblog contexts (3). It is therefore a strong classifier for our task, since Reddit posts are often short and we have over 2 billion comments to label. We used VADER's Python package *vaderSentiment* version 3.2.2 in this study. For the subreddit-level analysis, we first obtained the valence of comments, then averaged the valence of all comments in a given month by subreddit. Later, in the analysis in Section 1.2.4, comment valence was aggregated by week. For the thread-level analysis, we again measured every comment, but no aggregation was applied.

1.1.6. Measuring Semantic Diversity and Semantic Uniqueness

To obtain subreddits' monthly semantic uniqueness and comments' semantic diversity in threads, we used a deep learning model called *all-MiniLM-L6-v2* from HuggingFace to encode comments into 384-dimensional dense vectors. *all-MiniLM-L6-v2* is a sentence transformer, a specifically fine-tuned BERT model suitable for semantic textual similarity (STS) tasks (4). The model was set to encode up to 256 tokens, which is sufficient for most comments.

For thread-level analysis of semantic uniqueness, we similarly averaged the embeddings of the first 50 comments in a thread to obtain the semantic centroid. Then, for each comment, we calculated its cosine distance to this semantic centroid, which translates to its semantic uniqueness. High semantic uniqueness suggests that a comment is more different from typical comments in the thread, whereas low semantic uniqueness suggests that a comment is more similar to typical comments in the thread.

For subreddit-level analysis of semantic diversity, we averaged the embeddings of all comments in a given month by subreddit to obtain the semantic centroid. Then, for each comment, we calculated its cosine distance to this semantic centroid. Averaging all comments' cosine distances to the centroid yields the subreddit's semantic diversity for that month. High semantic diversity suggests that comments in that month are more different from one another, whereas low semantic diversity suggests that comments are more similar to one another. Later, in the analysis in Section 1.2.4, semantic centroid and semantic diversity were calculated on a weekly basis.

Note that both semantic uniqueness and semantic diversity are based on cosine distance between comments. They should therefore be interpreted with caution as indicators of differentiation. It may be the case that a user does not intend to differentiate, yet their posts are naturally more distinct because they introduce new topics to the discussion. In other words, semantic distance can be confounded by the range of topics available in the focal group. For example, a subreddit with high semantic diversity in a given month may reflect not only a stronger intention to differentiate but also a greater variety of topics available for discussion at that time. This is one reason we conducted thread-level analysis in addition to community-level analysis, which is highly aggregated. In threads, the submission constrains the topic of discussion, making high semantic uniqueness of a comment more likely to reflect a user's intention. Nevertheless, it is not a direct measure of intention to differentiate. Our experiment complements this by directly manipulating participants' intention.

1.2. Supplementary Observational Analyses

1.2.1. Environment and Regression Packages

Unless otherwise specified, all regression-related analyses were conducted in R version 4.4.2 using RStudio (MacOS version). Our mixed-effects regression models were performed using *lme4* version 1.1-36 and *lmerTest* version 3.1-3. The same software and packages were used for the experimental analyses.

1.2.2. Comparing the Valence of Old and New Users

We first developed an approach to identify new and old users for subreddits. For a given subreddit, we define new users as those who have never commented in the subreddit before, and old users as those who have commented previously. This definition involves a temporal element. For example, according to our definition, if a user joined r/democrats in March 2016, the user is considered a new user in that month but an old user thereafter. The definition is subreddit-specific: a new user in one subreddit may be an old user in another.

Using this approach and the corpora's metadata, we identified which comments were from new users and which were from old users. Since we had already determined every comment's valence with VADER, we were able to aggregate the valence of new and old users by month and examine the impact of cohort replacement on subreddit valence change. We also collected other statistics, such as the monthly counts of active new and old users, to include in our mixed-effects regression models.

When examining how the valence of old and new users changed over time, we used a within-between decomposition approach to separate (1) subreddits' average monthly user age relative to other subreddits and (2) subreddits' monthly age change within themselves. The latter was our main independent variable of time. When analyzing the valence of old users, we included the valence of new users in the model specification as a control variable, and vice versa. See **Table S1** and **Table S2** for the full results of the models reported in the main text.

Table S1.
Within-Subreddit Old Users' Valence Change

Predictor	*b*	SE	t value	p value	95% CIs
Old Users Valence					
Subreddit Age (within)	-0.04	0.003	-15.44	<0.001	-0.05, -0.04
Subreddit Age (between)	-0.09	0.01	-7.50	<0.001	-0.11, -0.06
New Users Count	-0.003	0.002	-1.25	0.21	-0.008, 0.002
Old Users Count	-0.01	0.003	-4.63	<0.001	-0.019, -0.008
New Users Valence	0.23	0.002	97.48	<0.001	0.23, 0.24

Note. Subreddits and Months are included as random effects.

Table S2.
Within-Subreddit New Users' Valence Change

Predictor	*b*	SE	t value	p value	95% CIs
New Users Valence					
Subreddit Age (within)	-0.02	0.003	-7.92	<0.001	-0.03, -0.02
Subreddit Age (between)	-0.11	0.01	-9.25	<0.001	-0.13, -0.09
New Users Count	-0.009	0.002	-4.14	<0.001	-0.013, -0.005
Old Users Count	-0.005	0.003	-1.85	0.06	-0.010, 0.0003
Old Users Valence	0.19	0.002	96.28	<0.001	0.19, 0.20

Note. Subreddits and Months are included as random effects.

1.2.3. Comparing the Valence of Reddit and News

News Articles Data Collection and Pre-processing. We used news articles from the News on the Web (NOW) corpus to compare against Reddit comments. The NOW dataset is a comprehensive collection of news articles scraped from a wide range of digitized newspapers and magazines, including not only traditional outlets such as *The Wall Street Journal*, *The Washington Post* etc., but also web-based news platforms such as Yahoo News, Fox News, CNN, etc.

The dataset is continuously updated, with the newest articles integrated into the corpus. The version we purchased and used included articles published between January 2010 and July 2022, allowing for direct comparison with the Reddit dataset. To focus on U.S. news content, we excluded non-U.S. articles, resulting in a sample of 11,688,074 articles. The distribution of articles varies over time, with a notable spike in 2021, when 6,229,686 articles were published. Earlier years in the dataset contain significantly fewer articles, likely due to a combination of more limited digital news production, restricted scraping access, and selective inclusion criteria during those years.

There are 22,661 unique publishers in the dataset. However, many correspond to the same publisher under slightly different names. To ensure consistency in analysis, we applied a series of preprocessing steps to standardize and match publisher names. As a result, we consolidated the dataset to 648 unique publishers. For further refinement, the dataset was deduplicated based on unique article IDs, reducing the total number of articles to 4,906,472. This deduplicated dataset serves as the primary source for all subsequent analyses.

To make the analysis parallel to subreddit partisanship ratings, we obtained publishers' partisanship bias from the AllSides Media Bias Rating, a public benefit corporation that uses blind bias surveys and editorial reviews to measure the perceived political bias of news outlets. The version we used was downloaded from here and corresponds to version 1.1, published in 2019. Some publishers' media bias ratings have changed in later versions, but we selected this version because it aligns with the time period of the articles we analyzed. Examining the subset of articles from publishers identified by AllSides suggested that media bias was tangential to article valence, $b = 0.02$, $SE = 0.03$, $t = 0.80$, $p = 0.43$, 95% CIs [−0.03, 0.07]. Therefore, we did not include source partisanship bias in our models, which allowed us to retain a larger sample for analysis.

When comparing the valence of Reddit and news, we used the valence of news article titles. We did not measure the valence of full articles for two reasons. First, due to dataset limitations, in the NOW corpus 10 words are randomly replaced by "@" in every 200 words to avoid copyright issues. This prevents accurate measurement of article-level valence. Second, since we used VADER to measure the valence of Reddit posts, methodological consistency required us to also apply VADER to the news data.

However, VADER is designed for short texts such as microblogs, not long-form news articles.

Using news article titles offers several advantages. Titles are often short, making them well-suited for VADER. Furthermore, prior research (5) suggests that negativity drives news consumption, and news outlets often frame article titles negatively to attract attention. This makes it unlikely that outlets systematically oversample positive events or frame negative events with positive titles. Thus, if we find that article titles have not become more negative over time, this provides stronger evidence against the claim that society has become more negative in the past decade.

One limitation of using news article titles is that some long titles are truncated in the NOW dataset, with approximately 25% ending in "…". This limits measurement accuracy. However, we did not observe systematic changes in the rate of truncation over time. Whether or not truncated titles were included, regression outcomes remained stable. In our subsequent analysis, we report results excluding truncated titles.

Regression Outcomes. We did find evidence supporting the alternative explanation that society has become more negative over the last decade. We aggregated news article titles' valence by publisher–month pairs and combined this dataset with the Reddit dataset, which was aggregated by subreddit–month pairs. We then fitted a model in which we interacted the time fixed effect with the corpus and included source of publication (i.e., news publishers or subreddits) and months as random effects. This model yielded a significant interaction effect, $b = -0.13$, $SE = 0.004$, $t = -29.50$, $p < 0.001$, 95% CIs [−0.14, −0.12], such that valence became more positive between 2010 and 2022 across news publishers, $b = 0.10$, $SE = 0.005$, $t = 19.60$, $p < 0.001$, 95% CIs [0.09, 0.11], but more negative across subreddits, $b = -0.02$, $SE = 0.004$, $t = -6.35$, $p < 0.001$, 95% CIs [−0.03, −0.02]. **Fig. S1** visualizes these trends.

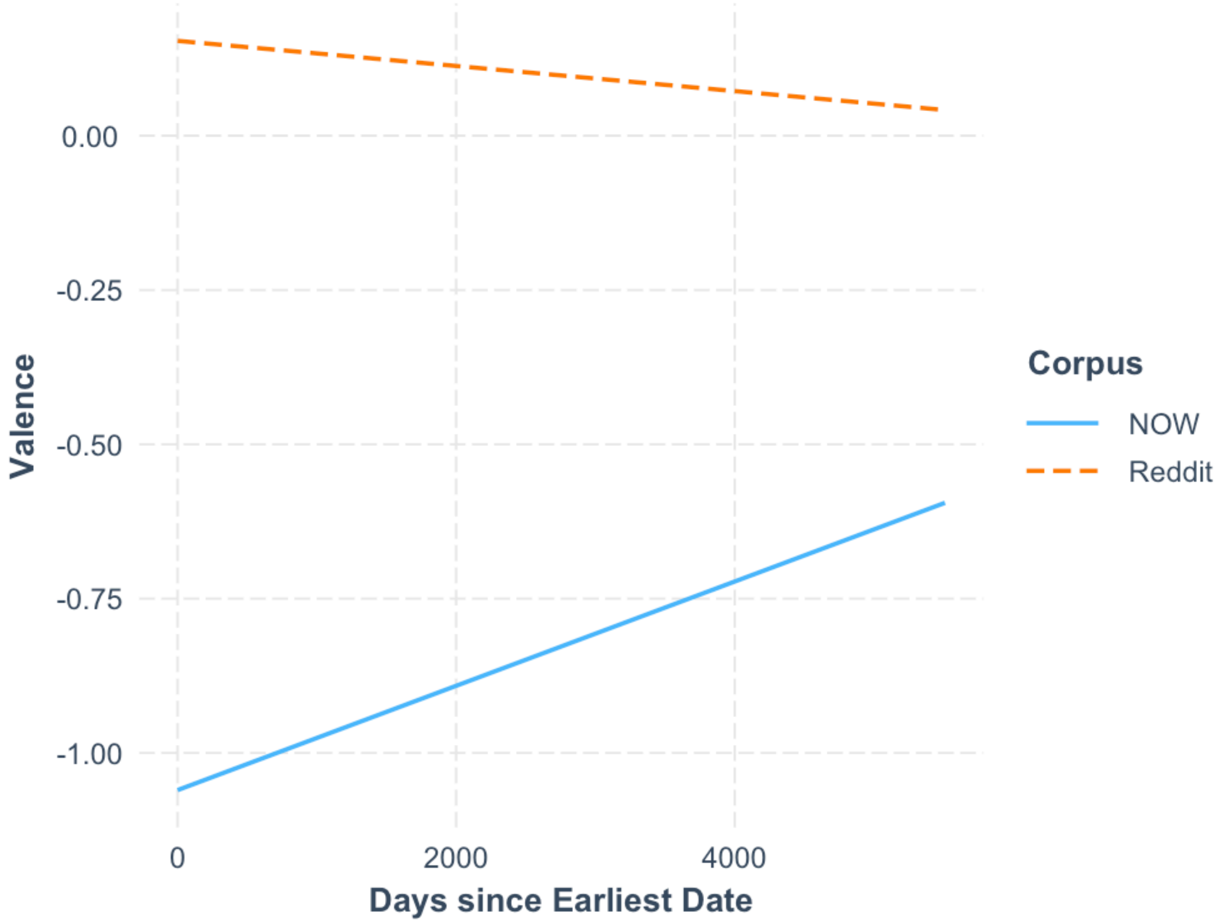

Fig. S1. Valence change of news articles and Reddit posts over time. The earliest date corresponds to the first appearance of a source's content in our dataset. For subreddits, this is their month of creation; for news publishers, it is the first month they were collected in the NOW corpus.

1.2.4. Moderation by Thread Initial Valence

There are two points that warrant clarification regarding how thread initial valence was operationalized. In this study, we define thread initial valence as the valence of the first comment posted in a thread. However, because many first comments are auto-generated bot posts, we uniformly removed the first comment in every thread and used the valence of the second comment to represent the thread's initial valence. Second, since only one comment was used to represent thread initial valence—and users may freely express opinions that do not necessarily align with the topic introduced by the original poster—this operationalization may be unstable and may not reliably capture the valence of a thread at its start.

To address these two potential limitations, we conducted an additional robustness check by including more early comments in a thread and using their aggregated valence to represent thread initial valence. In this section, we report the moderation effect of thread initial valence on the valence-declining effect when defining initial valence as the aggregated valence of the first 3%, 5%, or 10% of comments. Because every thread was uniformly truncated to the first 50 comments, the first 3%, 5%, and 10% correspond to the first 2, 3, and 5 comments, respectively (using a rounding-up strategy). As shown in **Tables S3–S5**, while the effect sizes vary slightly, they consistently resemble the moderation effect reported in the main text.

Table S3.
Moderation of Thread Initial Valence (first 3% comments)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Sequence	-0.003	0.0004	-7.11	<0.001	-0.004, -0.002
Thread Initial Valence	0.127	0.001	155.42	<0.001	0.126, 0.129
Sequence: Thread Initial Valence	-0.053	0.0004	-131.40	<0.001	-0.054, -0.052

Note. Subreddits and threads are included as nested random effects.

Table S4.
Moderation of Thread Initial Valence (first 5% comments)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Sequence	-0.003	0.0004	-7.11	<0.001	-0.004, -0.002
Thread Initial Valence	0.152	0.001	189.76	<0.001	0.151, 0.154
Sequence: Thread Initial Valence	-0.060	0.0004	-149.84	<0.001	-0.061, -0.059

Note. Subreddits and threads are included as nested random effects.

Table S5.
Moderation of Thread Initial Valence (first 10% comments)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Sequence	-0.003	0.0004	-7.16	<0.001	-0.004, -0.002
Thread Initial Valence	0.189	0.001	246.30	<0.001	0.187, 0.190

Sequence: Thread Initial Valence	-0.068	0.0004	-170.08	<0.001	-0.069, -0.068

Note. Subreddits and threads are included as nested random effects.

1.2.5. Moderation by Subreddit Initial Valence

Similar to thread initial valence, subreddit initial valence was defined as the aggregated valence of comments posted in the first month of each subreddit. While auto-generated bot comments are less of a concern for aggregated community-level analysis, this measure still carries the potential limitation of instability. Another potential issue is that only a few active users may contribute during the first month, making the comments less representative.

Parallel to the robustness check above, we included additional months to calculate subreddit initial valence. In this section, we report the moderation effect of subreddit initial valence on the valence declining effect when defining initial valence as the aggregated valence of the first 3, 6, or 12 months of comments. As shown in **Tables S6–S8**, while the effects vary slightly, they consistently resemble the moderation effect reported in the main text.

Table S6.
Moderation of Subreddit Initial Valence (first 3 months)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Subreddit Age (within)	-0.03	0.003	-13.01	<0.001	-0.04, -0.03
Subreddit Age (between)	-0.08	0.01	-6.02	<0.001	-0.10, -0.05
Subreddit Initial Valence	0.46	0.01	33.70	<0.001	0.43, 0.49
Subreddit Age (within): Subreddit Initial Valence	-0.06	0.001	-42.04	<0.001	-0.06, -0.06

Note. Subreddits and months are included as random effects.

Table S7.
Moderation of Subreddit Initial Valence (first 6 months)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Subreddit Age (within)	-0.03	0.002	-13.64	<0.001	-0.04, -0.03

Subreddit Age (between)	-0.05	0.01	-5.14	<0.001	-0.8, -0.03
Subreddit Initial Valence	0.57	0.01	50.44	<0.001	0.5,5 0.59
Subreddit Age (within): Subreddit Initial Valence	-0.07	0.001	-42.04	<0.001	-0.07, -0.07

Note. Subreddits and months are included as random effects.

Table S8.
Moderation of Subreddit Initial Valence (first 12 months)

Predictor	*b*	SE	t value	p value	95% CIs
Valence					
Subreddit Age (within)	-0.03	0.002	-13.42	<0.001	-0.04, -0.03
Subreddit Age (between)	-0.02	0.01	-2.87	0.004	-0.04, -0.01
Subreddit Initial Valence	0.65	0.01	74.35	<0.001	0.64, 0.67
Subreddit Age (within): Subreddit Initial Valence	-0.07	0.001	-49.26	<0.001	-0.07, -0.07

Note. Subreddits and months are included as random effects.

1.2.6. Cross-Correlation Analysis

To examine temporal associations between community size and language features such as valence and semantic diversity, we computed their cross-correlation functions after pre-whitening. Cross-correlation quantifies the similarity between two time series as one is shifted in time relative to the other. A critical concern in time-series analysis, however, is that raw correlations can be autocorrelated. To address this, we employed pre-whitening to remove autocorrelation between community size and language features before computing cross-correlations.

In the main text, we reported pre-whitened cross-correlations with lags at the monthly level. Since a month is a relatively large time interval, this choice may not be appropriate for capturing cross-correlations in online conversational dynamics. In these figures, for example, a lag of −10 would mean that increases in the number of active users in January are associated with semantic diversity in November (10 months later). However, conversations on Reddit are more likely to remain active for a few days up to a few weeks.

Because many subreddits do not have sufficient activity at the daily level, we retested the correlations between community size and valence and semantic diversity at the weekly level. Again, we found that increases in community size were most strongly associated with more negatively valenced comments within the same week, $b = -0.05$, $SE = 0.001$, $t = -49.94$, $p < 0.001$, 95% CIs [−0.049, −0.045], and with greater semantic diversity within the same week, $b = 0.08$, $SE = 0.001$, $t = 71.97$, $p < 0.001$, 95% CIs [0.076, 0.080], compared to other weeks for the same subreddit (**Fig. S2**).

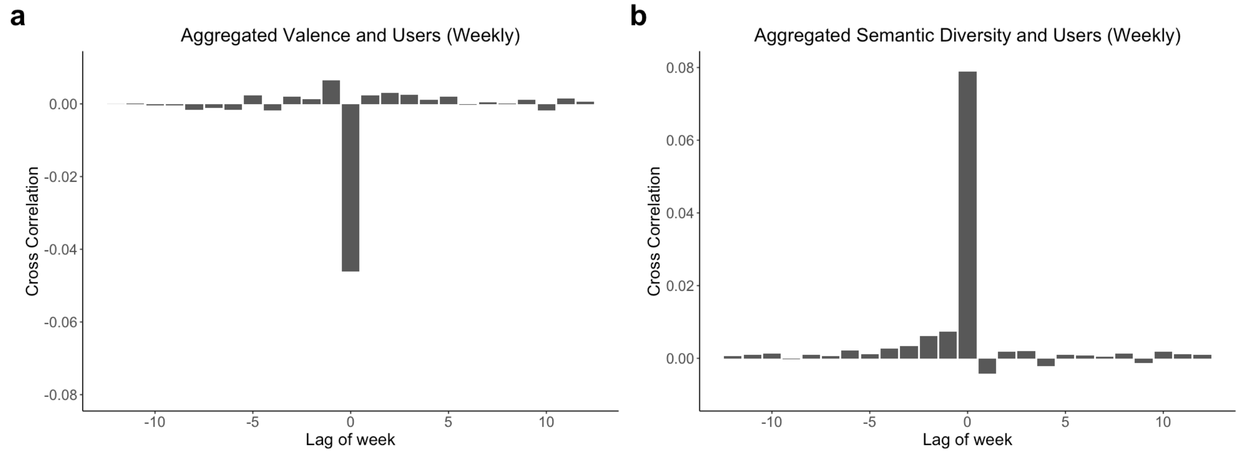

Fig. S2. Cross-correlations at weekly lags. a. Cross-correlation between valence and active user count. **b.** Cross-correlation between semantic diversity and active user count. **All** cross-correlations were pre-whitened to remove autocorrelation.

Further analysis of the contemporaneous cross-correlations suggests a possible non-linear effect of community size on semantic differentiation. In **Fig. S3**, we plotted subreddits' average weekly active users against their cross-correlations between community size and valence (**Fig. S3a**) and between community size and semantic diversity (**Fig. S3b**) at lag = 0. A key difference between the two figures is that there is a more distinct cluster of high cross-correlations between community size and semantic diversity among subreddits with low average weekly users. In other words, beyond a certain size, marginal increases in users no longer produce a noticeable difference in users' need for semantic differentiation.

On one hand, this may reflect a limitation of using semantic diversity as a proxy for differentiation, as discussed previously in Section 1.1.5. On the other hand, it may suggest that other factors that beyond the scope of our account drive the co-occurrence of increases in user numbers and decreases in valence in very large subreddits. For example, a controversial NBA game may lead many users to join the NBA subreddit to complain about a referee's decision. In this scenario, although their comments are largely negative, they may not necessarily differentiate themselves semantically from others. In sum, our conclusions should not be interpreted as universally applicable without constraints.

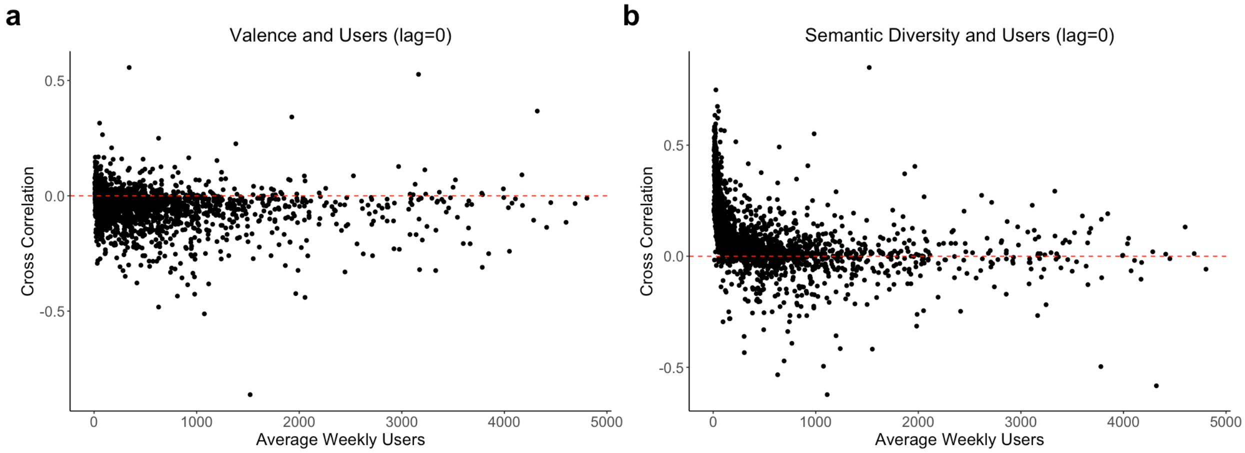

Fig. S3. Weekly cross-correlations at lag=0. a. Cross-correlation between valence and active user count at lag=0. **b.** Cross-correlation between semantic diversity and active user count at lag=0.

1.2.7. Moderation by the Proportion of Political Comments

Because our sample is non-random and tied to political engagement, it is important to examine how politics may moderate the valence decline effect we observed. As suggested by previous research (2,6), political engagement can make conversations more toxic, which may in turn make expressed valence more negative. Thus, there is a possibility that the valence decline effect was driven primarily by subreddits highly relevant to politics, but not applicable to others featuring general, non-political topics—thereby limiting the scope and theoretical implications of our findings. To test this, we explored how the proportion of political comments (derived from (2)) as an operationalization of political engagement moderates the valence-decline effect. This measure is closely tied to politics but is independent from the partisan score (derived from (1)) that we used to sample subreddits.

As shown in **Fig. S4**, we found that subreddits with higher levels of political engagement showed a slightly stronger valence-decline effect over time, $b = -0.01$, $SE = 0.001$, $t = -10.15$, $p < 0.001$, 95% CIs [−0.02, −0.01]. Valence in subreddits with a high level of political engagement (1 SD above the mean) declined rapidly, $b = -0.04$, $SE = 0.004$, $t = -10.89$, $p < 0.001$, 95% CIs [−0.05, −0.03], whereas valence in subreddits with a low level of political engagement (1 SD below the mean) declined slightly in comparison, $b = -0.01$, $SE = 0.004$, $t = -3.00$, $p = 0.004$, 95% CIs [−0.02, −0.004]. This may imply fiercer competition for attention in politically oriented communities compared to non-political ones. Subreddits with lower levels of political engagement also displayed a weaker but still consistent valence-decline effect over time. In our sample, the mean proportion of political comments across subreddits was 11.4%, which is below the 25% threshold used in (2) to define political subreddits. Therefore, we are confident that the valence-decline effect is applicable to communities with varying levels of political engagement.

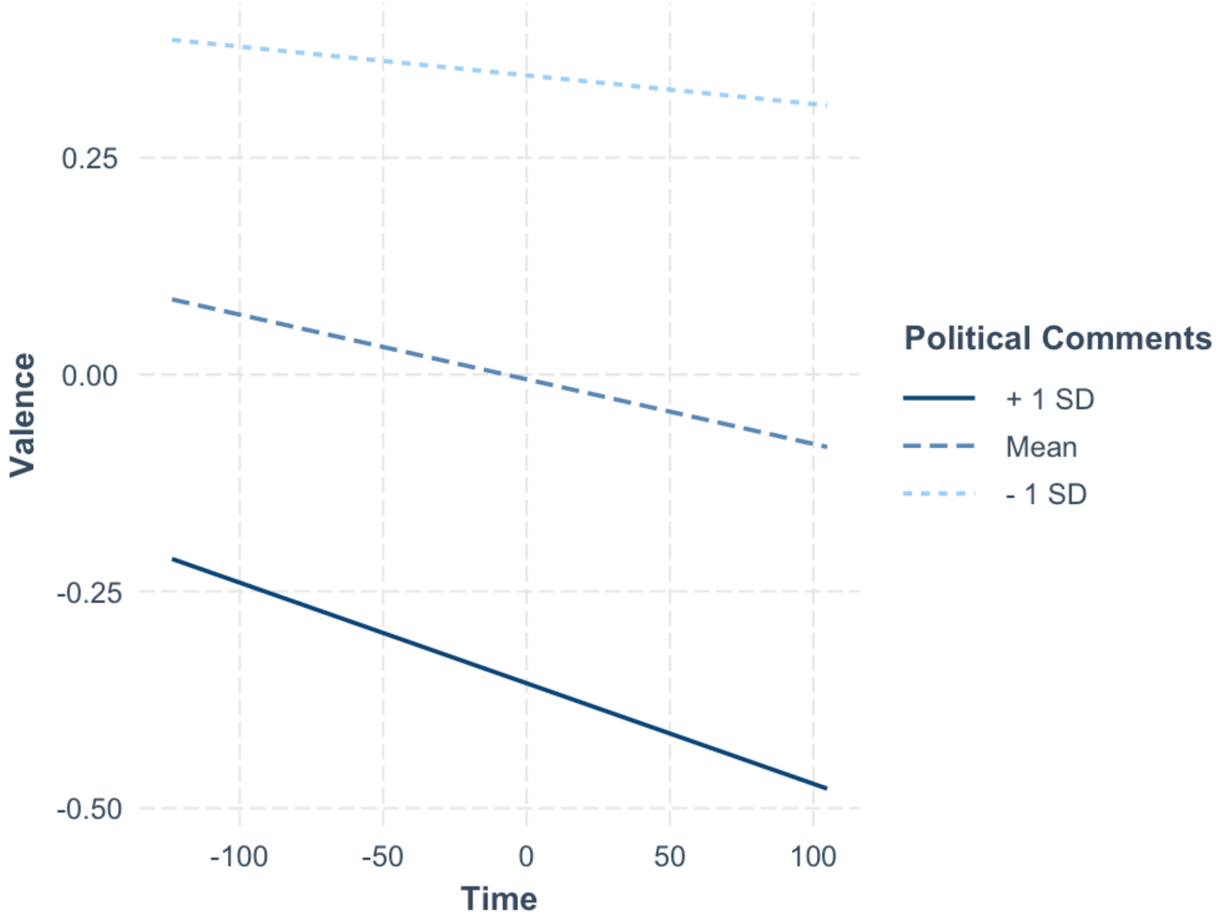

Fig. S4. Political comments proportion as a moderator of the valence decline effect over time in subreddits. Subreddits with a higher proportion of political comments show lower valence and a faster rate of decline, whereas subreddits with a lower proportion of political comments show higher valence and a slower rate of decline.

1.2.8. Valence, Semantic Uniqueness, and Comment Score

This section tests whether semantic differentiation (measured in the way we report in the main text) is associated with a higher comment score. One limitation of this analysis is that Reddit only publishes each comment's total score (the sum of upvotes and downvotes), meaning that comments that elicit very low engagement (e.g., 1 upvote and 1 downvote) could hypothetically receive the same aggregate score as comments that elicit very high engagement (e.g., 50 upvotes and 50 downvotes). Nevertheless, analyzing aggregate scores gave us preliminary insight into whether unique content was rewarded with more popularity. We selected comments with scores ranging from 0 to 500, which account for approximately 96% of all comments, and operationalized comment popularity as the logged score.

Preliminary plots suggested that semantic differentiation was quadratically associated with score, with the moderately unique comments receiving the highest score (see **Fig. S5**). To formally test this association, we fit a mixed effect regression model at the thread level in which the log of score was regressed on linear and quadratic semantic differentiation terms (both centered within threads), a between-thread semantic differentiation term, within- and between-thread valence terms, and terms controlling for comment sequence and level. The regression nested comments within threads, which were in turn nested in subreddits.

In this regression, the linear term was significant and negative, $b = -0.56$, $SE = 0.01$, $t = -220.96$, $p < 0.001$, 95% CIs [−0.56, −0.55], and the quadratic term was significant and positive, $b = 0.62$, $SE = 0.01$, $t = 48.72$, $p < 0.001$, 95% CIs [−0.60, −0.64]. As in **Fig. S5,** moderately unique content appeared to receive the highest overall score.

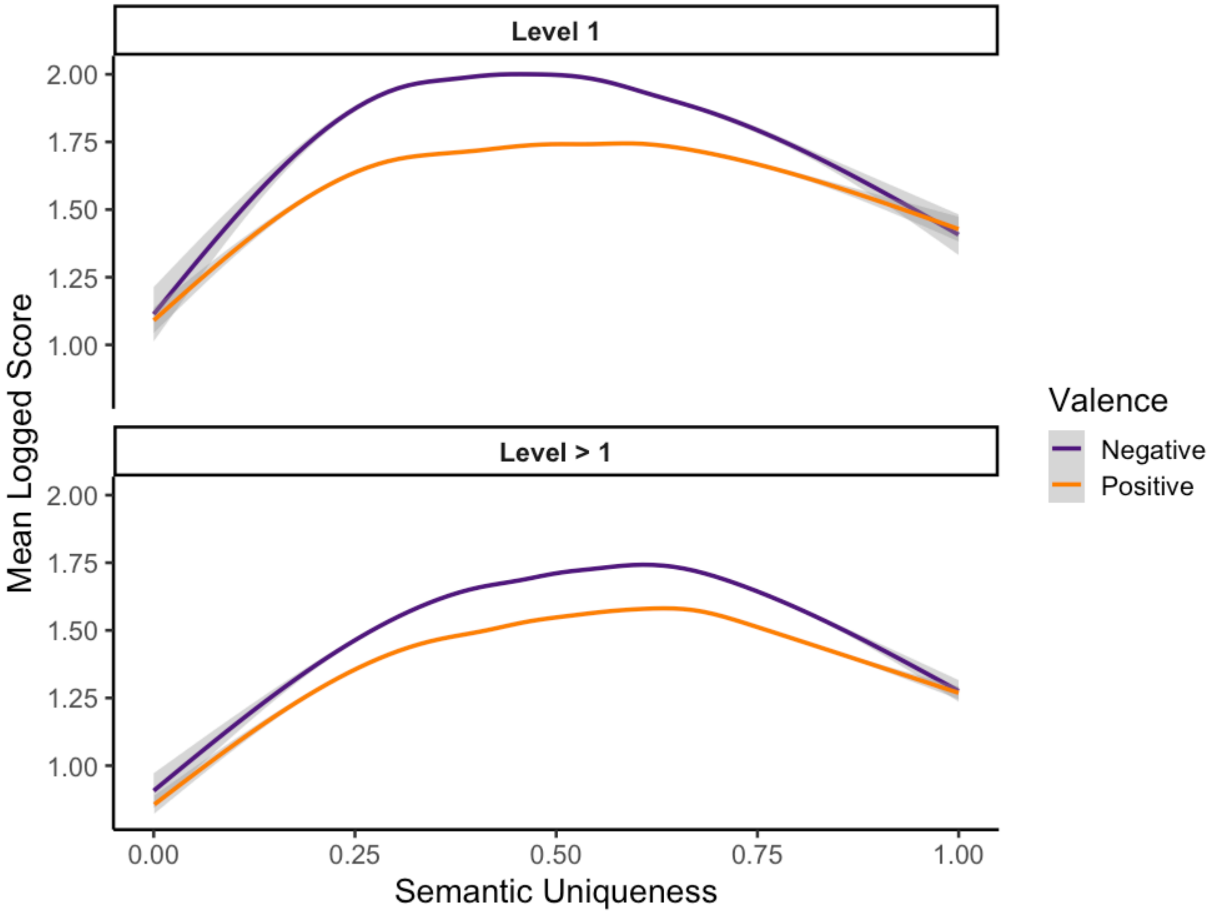

Fig. S5. Post semantic uniqueness and popularity. The upper panel shows the association between comment semantic uniqueness and mean logged score for comments that directly reply to the original post. The lower panel shows the association for comments that reply to other comments. Lines were fitted using GAM.

Interestingly, at the between-thread level, semantic differentiation was robustly and positively associated with score, *b* = 1.83, *SE* = 0.02, *t* = 82.14, *p* < 0.001, 95% CIs [1.78, 1.87], suggesting that more semantically differentiated threads received higher scores in general. The regression also revealed that valence was negatively associated with comment score within threads, *b* = -0.03, *SE* = 0.0007, *t* = -44.53, *p* < 0.001, 95% CIs [−0.03, −0.03], and between threads, *b* = -0.35, *SE* = 0.009, *t* = -40.44, *p* < 0.001, 95% CIs [−0.37, −0.34], indicating that negative comments received higher scores. This pattern, illustrated in **Fig. S6,** is consistent with past work showing that negativity drives engagement on social media (7).

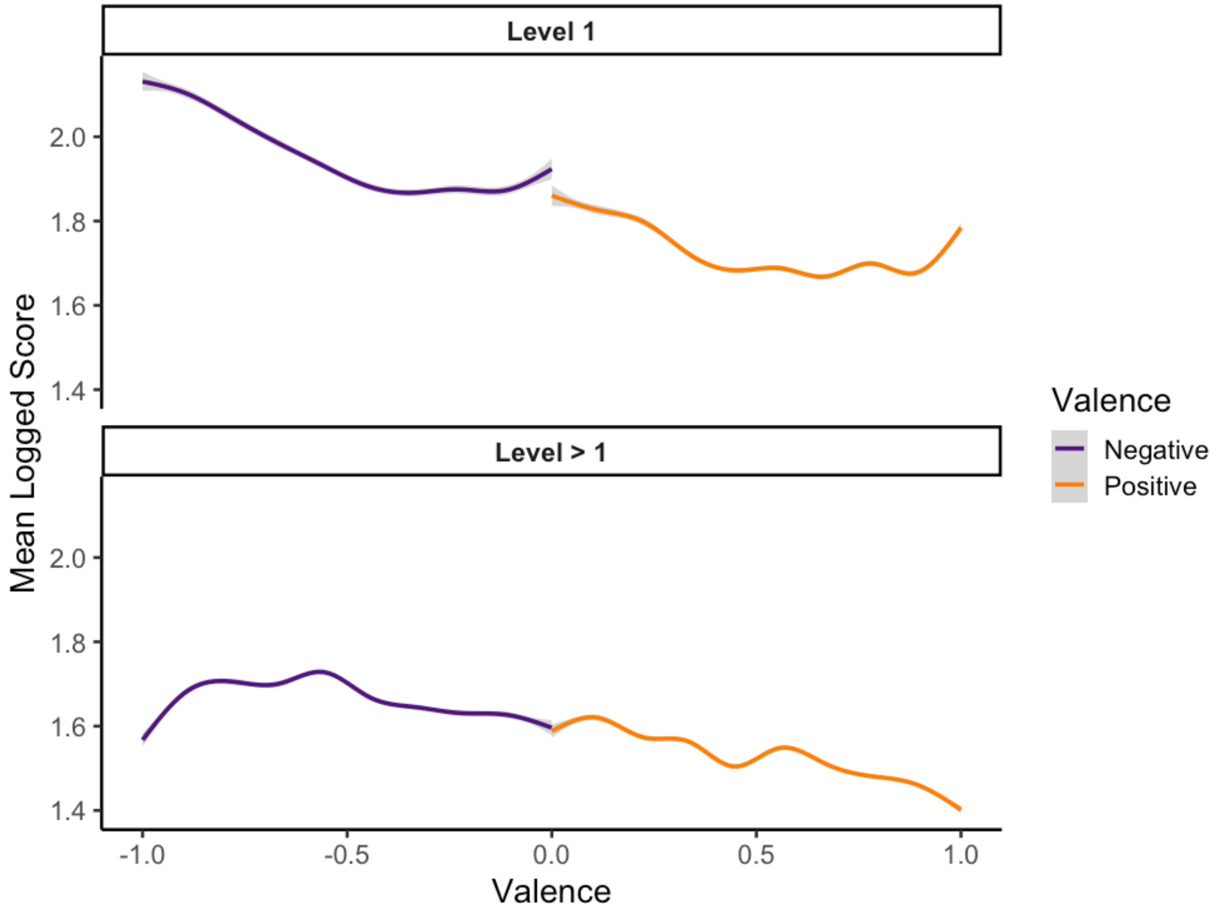

Fig. S6. Comment valence and popularity. The upper panel shows the association between comment valence and mean logged score for comments that directly reply to the original post. The lower panel shows the association for comments that reply to other comments. Lines were fitted using GAM.

We view this research on incentivized negativity as complementary to our own findings. Users may independently post negative comments because they are directly incentivized to do so (past work), but also as a byproduct of trying to differentiate themselves (the present work). Our analyses raise an interesting question about whether uniqueness does indeed drive engagement. While our findings are mixed, this may be because aggregate score is a coarse measure of engagement–failing to identify posts that were engaging and controversial.

Another key limitation of this approach is that we do not know the contexts of the comments. There are many reasons why a comment may be negative: a user could be expressing moral outrage or disagreeing with others, and such comments may attract upvotes for different reasons. We also lack a clear understanding of the subreddits in which these comments were posted, which may influence user behavior. For example, in some subreddits it may be normative to upvote more frequently, while in others it may be less common. Therefore, while we claim that there are associations between comment popularity and comment valence and semantic uniqueness at the aggregated level, we do not have a clear answer to whether negative comments are consistently rewarded more than others on Reddit.

1.2.9. Temporal Change of the Valence Decline Pattern

Since our dataset incorporates Reddit comments over a decade, one may ask whether threads from 10 years ago exhibit the same valence decline trend as those more recent. Over this period, many factors have changed, such as website design, user demographics, and active communities, all of which could influence commenting behavior. It is therefore natural to ask whether the observed valence-decline trend persists over time.

We examined how time moderates valence-decline trends in threads on a yearly basis. The year a thread was created was determined by the creation time of its first comment. We found that older threads exhibit a stronger valence-decline trend compared to more recent threads (b = 0.002, SE = 0.0004, t = 5.05, p < 0.001, 95% CIs [0.001, 0.003]. The trends are visualized in **Fig. S7**. This result differs from our expectation that the valence decline trend would be stronger in recent years. One possible explanation is that in recent years, comments have been more negative overall, b = −0.006, SE = 0.001, t = −5.77, p < 0.001, 95% CIs [−0.008, −0.004], leaving less room for valence to decline within threads.

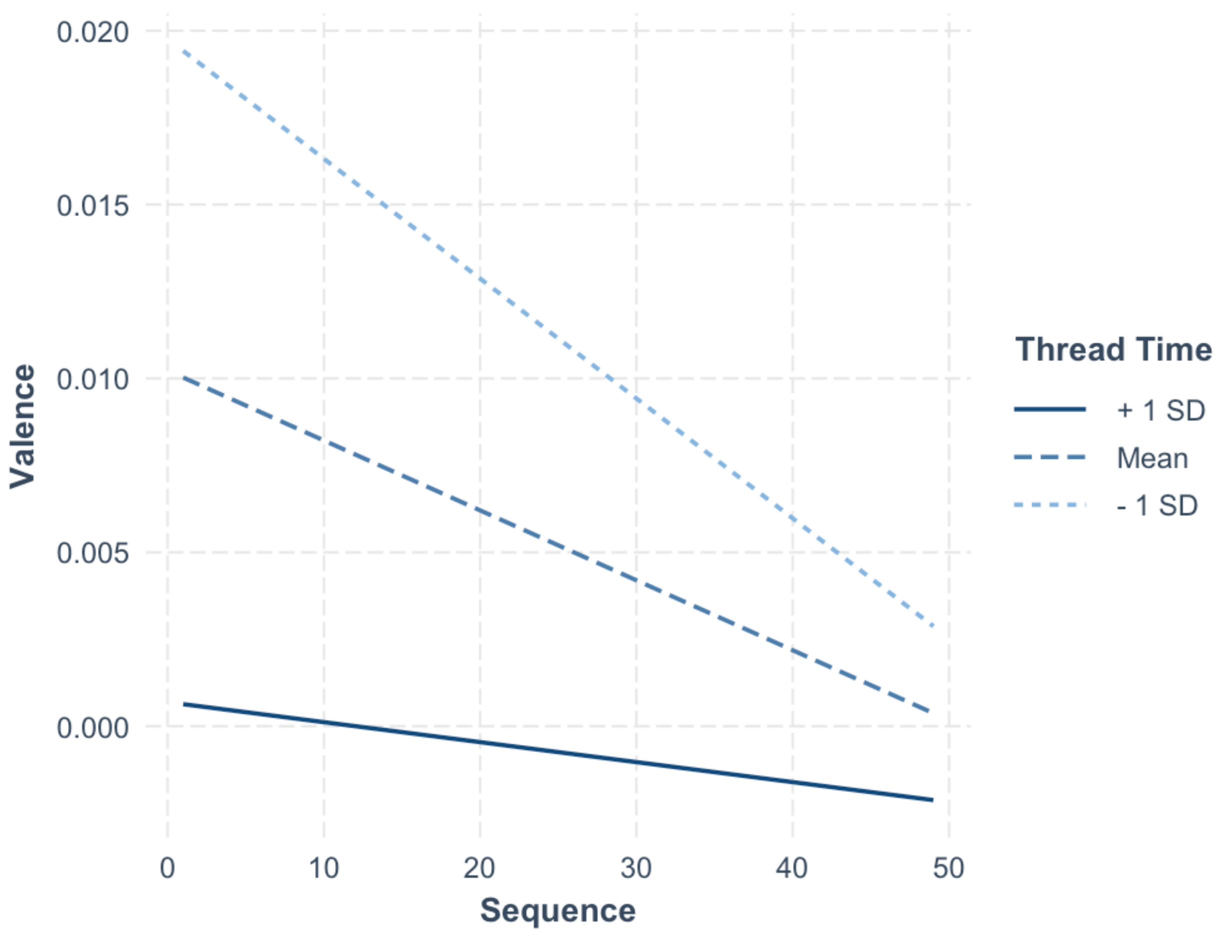

Fig. S7. Valence decline in threads moderated by thread creation time. Threads created in earlier years show steeper valence decline trends compared to those in recent years. Comments in threads created in recent years are overall more negative than those in earlier years.

2. Experiment

2.1. Experimental Design

2.1.1. Power Analysis

We determined the sample size for the experiment reported in the main text using a power analysis with the *simr* package (with publicly available code hosted here). This power analysis was seeded with an effect size from our supplemental experiment (see Section 2.3), which we conducted prior to the main-text experiment. In that supplemental experiment, we detected an interaction between participant generation and initial valence that supported our theory: when participants discussed positive news articles, their discussions became more negative over time, whereas when they discussed negative news articles, their discussions became more positive over time.

We made two changes in our subsequent experiment, both of which were incorporated into our power analysis. The first change was that participants were assigned to discuss either positively or negatively valenced news stories, rather than both within the same experimental session. In other words, we manipulated valence between subjects rather than within subjects. To implement this in the power analysis, we based our effect size on a truncated version of the original dataset, in which we randomly sampled data from either the positive or negative article conditions for each participant. The second change was the introduction of social incentives to either encourage conformity or incentivize differentiation (uniqueness). Thus, we needed to sample participants with the goal of achieving sufficient power to detect a three-way interaction between generation, article valence condition, and incentive condition.

The effect size of the interaction term from the between-subjects version of our previous experiment was r = 0.074. We seeded the power analysis with this term, assuming the same mixed-effects model specifications as in our previous experiment, and ran 1,000 simulations to calculate power at an alpha level of 0.05. We re-calculated these simulations for sample sizes ranging from n = 200 per incentive condition to n = 2000 per incentive condition, in intervals of 200 (i.e., 200 participants, 400 participants, 600 participants, etc.). We used the PowerCurve function from *simr* to calculate and visualize these simulations. The results suggested that a sample size of 800 participants per condition (1,600 total) would be sufficient to detect our hypothesized two-way interaction at 80% power, and 1,000 per condition (2,000 total) would be sufficient to detect the interaction at 90% power (**Fig. S8**).

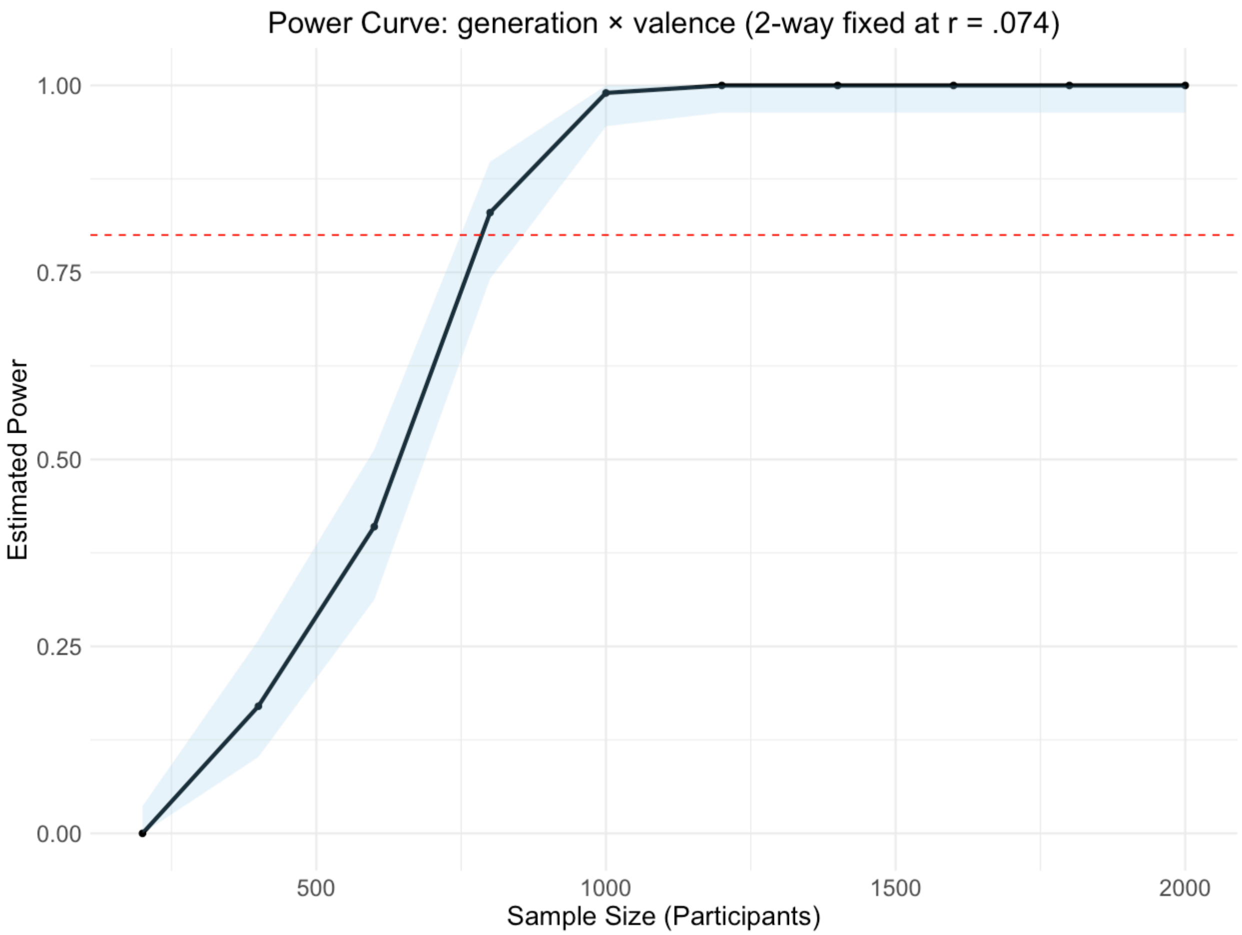

Fig. S8. Power curve for detecting a two-way interaction between article valence and generation. The dashed red line indicates the 0.80 power threshold.

We used a similar simulation process, anticipating a small effect size (*r* = 0.10) for our hypothesized three-way interaction between generation, valence condition, and incentive condition. The results suggested that a sample size of 800 participants per condition (1,600 total) would be sufficient to detect the hypothesized three-way interaction at 80% power, and 1,000 per condition (2,000 total) would be sufficient to detect the interaction at 90% power (**Fig. S9**).

Ultimately, we decided to double the 90% power sample size—recruiting 2,000 participants per condition (4,000 total)—to increase the likelihood of being sufficiently powered to observe the hypothesized three-way interaction. As reported in the main text, we did indeed find this three-way interaction.

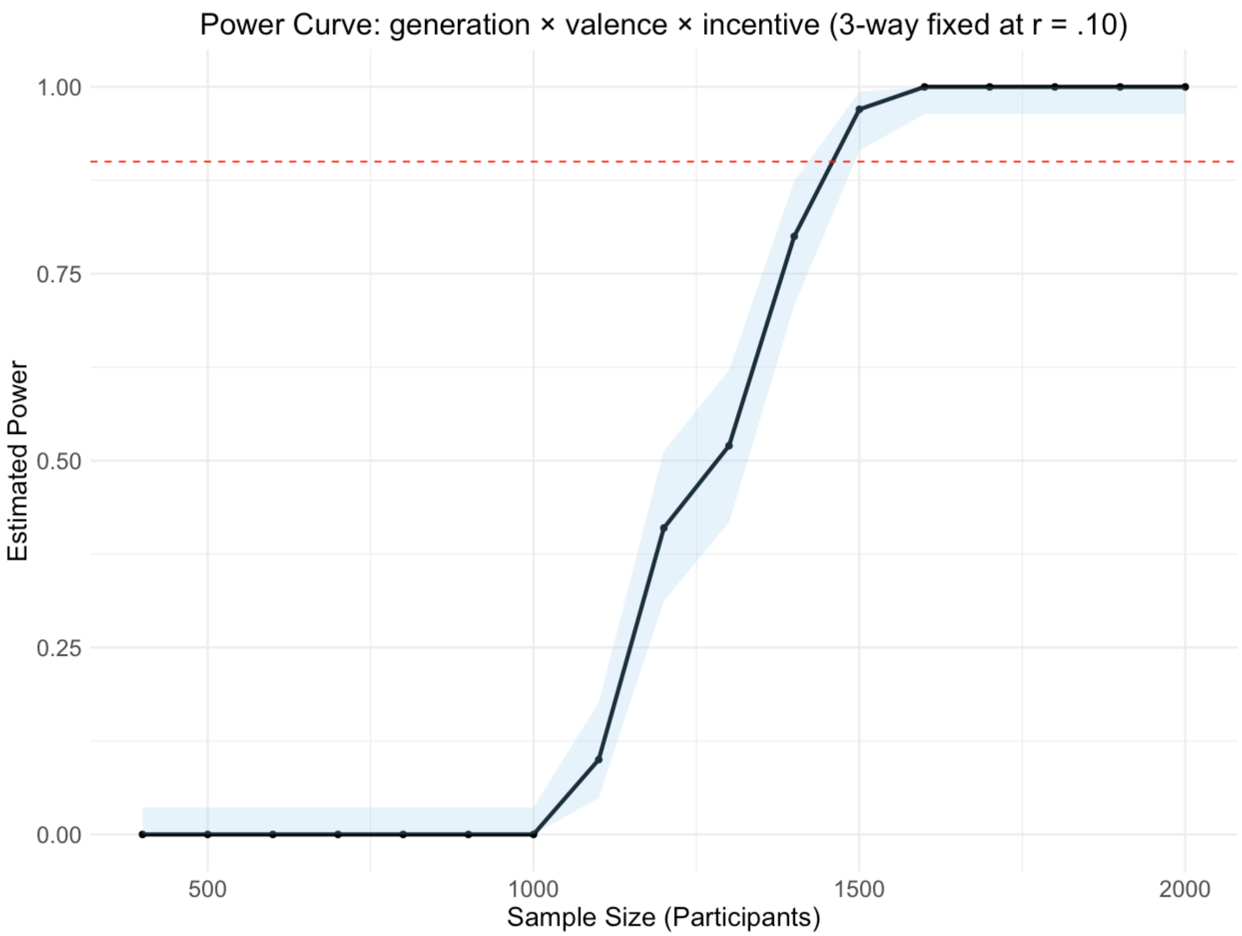

Fig. S9. Power curve for detecting a two-way interaction between article valence and generation, and incentive. The dashed red line indicates the 0.80 power threshold.

2.1.2. News Selection

For the main experiment, we scraped and selected real news headlines from Google News to display to participants. On October 11, 2024, we collected approximately 200 news articles from the *Top Stories* section of the website. We then used the GPT-4o API to assess the headlines on three dimensions: impact, valence, and partisanship (see Section 3.1 for the prompts). Two researchers, including the first author, manually reviewed the headlines and their GPT-4 ratings. Six headlines were initially selected based on the criteria that they had similar impact and partisanship ratings but differed in valence ratings. Two were later excluded after a pilot survey revealed that one was still related to politics and the other was difficult to comprehend. The four headlines selected for the formal generational experiment were:

1. (Positive) A Japanese Organization of Atomic Bombing Survivors Wins the 2024 Nobel Peace Prize
2. (Neutral) Hundred-Year-Old Remains Of British Man Who Died On Everest Discovered By Nat Geo Doc Team Including, Free Solo’s Jimmy Chin
3. (Moderately negative) Sealed TikTok court documents show time limit tool effectively did nothing to reduce teen usage, NPR reports
4. (Highly negative) Hurricane Milton Live Updates: Death Toll Climbs as Florida Assesses Storm Damage

In the third headline, the original ending “NPR reports” was replaced with “according to new reports” for use in the experiment.

2.1.3. Experimental Procedure

Seed Survey. Before conducting the five generational surveys, we ran a seed survey to collect initial responses to different news headlines. The purpose of this survey was to check whether the selected headlines were appropriate for the generational experiment and to gather initial responses that could be displayed to participants in that experiment.

In the seed survey, participants were asked to write responses to three headlines randomly selected from a pool of six. We recruited 200 participants, resulting in 600 responses across news headlines of varying valence. As noted previously, two of the headlines and their associated 200 responses were not used in the formal generational experiment. Participants were compensated $0.80 for their participation, in accordance with the standard rate on Prolific. We screened participants to ensure they were U.S. citizens and used English as their primary language.

Generational Surveys. Our five-generational experiment used a generation × valence condition × incentive condition between-subjects design. In each generation, participants read either two positively valenced news headlines (one of which was rated neutral by GPT-4) or two negatively valenced headlines, and they were incentivized to write either a unique or a conforming response.

The incentive manipulation was implemented by displaying participants a set of responses to the same headline and instructing them either to conform to the perspectives shown or to differentiate their response from them. The number of the responses is 20, and they were randomly sampled from the same valence and incentive condition in the previous generation. For the first generation, the 20 responses were drawn from the seed survey. Participant responses were cleaned using the GPT-4 API to correct typos and grammar issues before being used in the next generation survey, and low-quality responses (e.g., off-topic or irrelevant responses submitted to cheat for completion) were removed in this process. Importantly, if either of a participant's two responses was judged low-quality, both were removed from the sample. The language model prompt used for cleaning is provided in Section 3.2.

To improve compliance, we informed participants that the person who wrote the most unique response (in the differentiation condition) or the most similar response (in the conformity condition) would receive an additional $5 reward. All participants were also informed that they would only be eligible if their posts were written in their own words, addressed the event, and were not generated with ChatGPT or similar models. Uniqueness or similarity was operationalized as the cosine similarity between participants' responses and the responses displayed to them. In each generation, the participants whose responses were the most unique or most conforming (relative to all others) were awarded. In total, we selected 10 winners and distributed $50 in bonus payments after the experiment concluded. The actual survey prompt shown to participants is available at https://osf.io/68wvm/.

Because this was a multi-generational experiment, we screened participants so that no one could take part in multiple generations. This exclusion also applied to participants in the seed survey and the supplementary multi-generational experiment, since those surveys shared similar designs. We recruited 4,000 participants and collected 8,000 responses. Participants were compensated $0.75 for their participation. All participants were screened to ensure they were U.S. citizens and used English as their primary language.

2.2. Supplementary Experiment

2.2.1. Experiment Data Details

Responses. We expected to recruit 4,000 participants, with 200 participants writing 400 responses for each generation–valence–incentive condition, totaling 8,000 responses. In practice, the actual number of responses varied due to platform-related factors (e.g., concurrent submissions, timeouts) and our data-cleaning process. Below, we report the actual number of responses collected for analysis in each generation.

Table S9.
Collected responses in each condition

Gen	Differentiation Negative	Differentiation Positive	Conformist Negative	Conformist Positive
1	374	366	378	354
2	368	362	406	358
3	348	334	402	364

4	354	362	378	360
5	372	354	394	382
N	1,816	1,778	1,958	1,818
Total				7,370

Note: Each participant wrote two responses. Dividing the number of responses in each cell by two gives the number of participants, which was 3,685.

We noticed that the conformist–negative condition consistently yielded more qualified responses across generations compared to other conditions. Most of this difference arose from the cleaning process, which represents a potential limitation of our experimental analysis. On the other hand, it may also suggest that there are more diverse ways to be negative than to be positive, consistent with prior work (8,9). In the experiment, participants were explicitly instructed to write responses in their own words, regardless of condition. We reason that because negativity can be expressed in more diverse ways than positivity, it may have been easier for participants to write negative responses. As a result, there was less need for participants to generate low-effort, low-quality responses, which are most likely to be filtered out during cleaning.

Participant Demographics. Since this was a multi-generational experiment, we report participant demographics by generation. See **Table S10.** for details.

Table S10.

Participant Demographics

Variable		**Gen 1**	**Gen 2**	**Gen 3**	**Gen 4**	**Gen 5**
Age	Age	40.8 (11.8)	41.0 (12.6)	40.3 (12.5)	39.3 (12.5)	39.2 (12.1)
Gender	Male	30.4%	31.3%	34.9%	36.6%	33.7%
	Female	68.2%	66.5%	63.4%	61.8%	64.2%
	Other	1.4%	2.1%	1.7%	1.7%	2.1%
Education	No High School	0.5%	1.1%	1.1%	1.1%	0.8%
	Technical Training	0.4%	0%	0.8%	0.1%	0.1%
	High School	30.6%	29.9%	32.3%	29.2%	27.0%
	Associate	15.6%	15.8%	11.6%	13.3%	16.0%
	Undergraduate	33.8%	34.5%	36.7%	34.5%	34.2%
	Graduate	19.0%	18.7%	17.4%	21.7%	21.8%
Partisanship	Republican	22.0%	28.4%	26.1%	25.6%	27.2%
	Democrat	40.9%	38.8%	38.7%	40.3%	39.8%

Independent	31.9%	26.5%	30.2%	29.6%	28.2%
Other	1.6%	2.5%	2.2%	1.7%	2.0%
No Preference	3.5%	3.7%	2.8%	2.9%	2.8%

Note: We report mean values (with SD in parentheses) for age and percentages for the other variables. Participants who selected the age category “60 or older” were assigned an age value of 60. Some category totals may not sum to 1 due to rounding.

2.2.2. Manipulation Check

We checked the incentive manipulations for the differentiation and conformist conditions across generations. For each participant’s response, we calculated its cosine similarity with the 20 randomly selected comments displayed to them, then averaged these 20 similarity scores to obtain the response’s overall similarity to the reference set. For all responses to the same headline within a given generation–incentive condition, we averaged their similarity scores and plotted them (**Fig. S10**). As expected, responses in the conformist condition became increasingly similar to the displayed comments, whereas those in the differentiation condition remained more distinct. Consistently, similarity scores were lower in the differentiation condition than in the conformist condition, suggesting that our manipulation was effective.

One point to note is that similarity scores in the differentiation condition did not significantly change across generations. One reason is that, regardless of how varied some responses are, they are rarely perfectly unrelated in high-dimensional vector space—especially when considering large numbers of responses—which sets a limit on the lower bound of aggregated similarity. Another possible reason relates to the nature of differentiation within our multi-generational design. Because the topics were fixed, participants could use their imagination to some extent, but the actual framings available to make their comments unique were limited. As a result, participants may have circled around similar framing strategies across generations, causing similarity scores to fluctuate but remain stable overall.

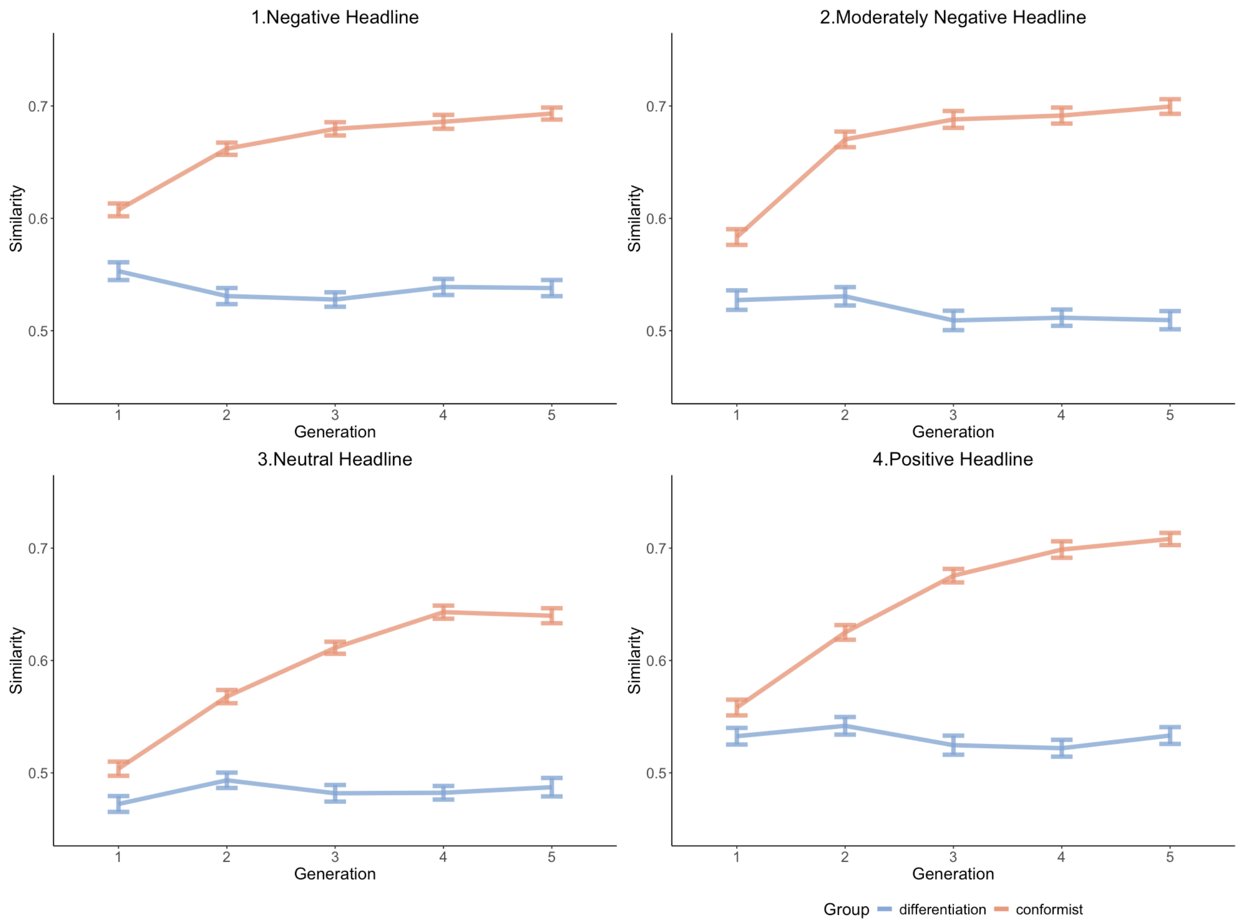

Fig. S10. Cosine similarity between participants' responses and reference comments across generations. Regardless of headline valence, responses in the conformist condition became more similar to the reference comments, while responses in the differentiation condition remained distinct. Similarity scores in the differentiation condition were consistently lower than those in the conformist condition.

2.2.3. Replication with VADER

We replicated the main experiment's major finding using participants' response valence measured by VADER. This was not a pre-registered analysis, but we conducted it as an exploratory test to be as transparent as possible. We fitted regressions with the same setup as in the main text, except that response valence was measured with VADER rather than GPT-4. We first regressed comment valence on generation, headline valence condition, incentive type, and the interaction terms of these fixed effects. In this regression, we again found evidence for our hypothesized three-way interaction, *b* = 0.03, *SE* = 0.01, *t* = 3.10, *p* = 0.002, 95% CIs [0.01, 0.04].

However, unlike when valence was measured with GPT-4, we were only able to replicate the positive interaction between valence and generation for participants incentivized to conform, *b* = 0.02, *SE* = 0.006, *t* = 2.96, *p* = 0.003, 95% CIs [0.01, 0.03]. We did not replicate the negative interaction between valence and generation for participants incentivized to differentiate, *b* = −0.01, *SE* = 0.006, *t* = −1.42, *p* = 0.15, 95% CIs [−0.02, 0.003].

As illustrated in **Fig. S11**, the reason we still observe a significant decline in valence for participants in the positive condition is that valence does not show changes over time in the other conditions. This may be because VADER is a contextually insensitive classifier (10, 11), and therefore inaccurately measured some participant responses. For instance, a comment such as “Oh great, another political stunt. Just what we need.” in response to political news would be classified by VADER as positive because of the word “great,” even though the overall meaning is clearly negative. In contrast, GPT-4 correctly identified the sarcastic tone and rated the response as negative. This suggests that our observational findings might have been even stronger with a more accurate measure of valence. However, GPT-4 is not feasible to scale to an analysis of 2 billion comments.

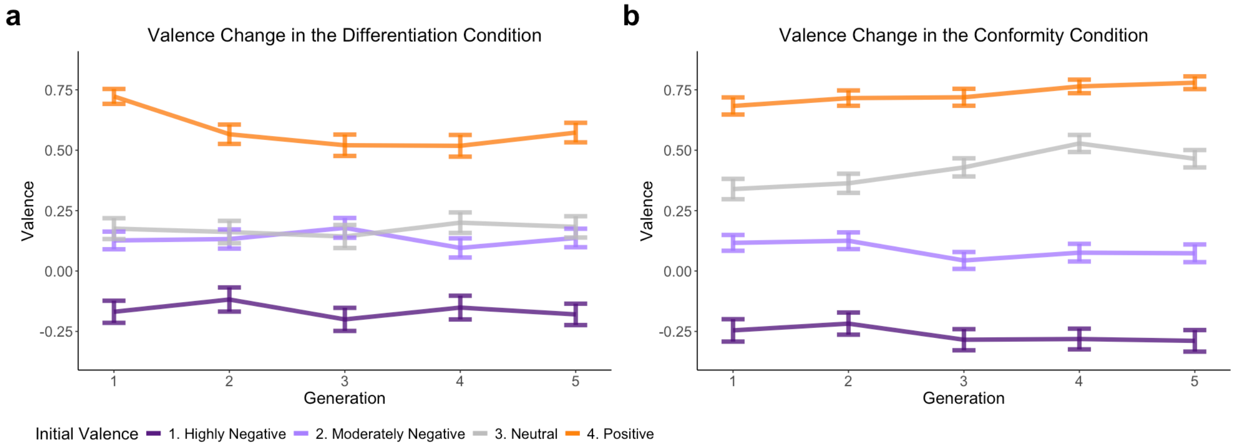

Fig. S11. Valence change across generations in conformity and differentiation conditions using VADER. a. Valence of participants’ responses to different news headlines across generations in the differentiation condition. **b.** Valence of participants’ responses to different news headlines across generations in the conformity condition.

2.3. Replication Experiment

2.3.1. News Selection

Two authors manually selected four news summaries from the Wikipedia page [2024 in the United States]() based on their valence, political relevance, and recency. We aimed to ensure that all the events had occurred recently by the time we released the surveys, while still differing in valence and political relevance. We included politics to test whether our effects varied depending on the salience of intergroup competition. The four news summaries were:

1. (Negative, political-irrelevant) Tornadoes strike Oklahoma, Texas, Arkansas, and Kentucky, killing 21 people
2. (Negative, political-relevant) In Planned Parenthood Arizona v. Mayes, the Arizona Supreme Court upholds an 1864 law that disallows most types of abortions
3. (Positive, political-irrelevant) Boeing's Starliner capsule launches its first astronaut-crewed flight into space to the International Space Station after several delays at the Cape Canaveral Space Force Station in Florida.
4. (Positive, political-relevant) The Biden administration released revisions to Title IX rules which give further protections for LGBTQ+ students as well as parenting and pregnant students.

One thing to note is that political relevance interacts with valence, such that a piece of news may be considered positive by one partisan group but not by those with different political orientations. Because our sample was skewed toward participants who self-identified as Democrats, we labeled the valence of politically relevant news summaries from a liberal perspective. In our demographic reports, we later confirmed that there were indeed more Democrats than others in the sample, and we showed that whether a news summary was politically relevant or not did not affect our conclusions.

Another point to note is that the Wikipedia page is continually maintained by community members, and some of the news summaries we used are no longer listed. Readers may refer to the page's editing history for a record.

2.3.2. Experimental Procedure

The experiment employed a mixed 5 (generation, between-subjects) × 2 (initial valence, positive vs. negative, within-subjects) design. In the first generation, participants were prompted to write four responses to the four news summaries described above (two positive, two negative). The order of the news summaries was randomized. For participants in subsequent generations, we additionally required them to read 25 comments on each news summary from earlier participants before writing their own responses. These responses were drawn from the pool of the previous generation and were cleaned for grammar using the GPT-4 API, during which low-quality responses were also flagged. If any one of a participant's four responses was judged to be low-quality, all of their responses were removed from the sample. The language model prompt is provided in Section 3.2, and the actual survey prompt shown to participants is available at https://osf.io/68wvm/.

Participants were compensated $0.75 for their participation. All participants were screened to ensure they were U.S. citizens and used English as their primary language, and no participant could take part in multiple generations.

2.3.3. Data Curation and Analyses

Responses. We expected to recruit 1,000 participants, with 200 participants writing 800 responses in each generation. In total, 1,039 participants completed our survey, with 199, 203, 203, 201, and 206 participants in each generation, respectively. After removing low-quality responses, the number of participants retained in our sample was 194, 201, 200, 199, and 206 across the five generations, totaling 1,000.

Participant Demographics. Since this was a multi-generational experiment, we report participant demographics by generation. See **Table S11.** for details.

Table S11.

Participant Demographics

Variable		Gen 1	Gen 2	Gen 3	Gen 4	Gen 5
Age	Age	35.7 (11.2)	36.0 (11.3)	36.3 (12.4)	36.6 (11.4)	35.8 (11.0)
Gender	Male	35.1%	38.8%	31.5%	28.6%	29.6%
	Female	61.9%	57.2%	66.5%	68.8%	66.0%
	Other	3.1%	4.0%	2.0%	2.5%	4.4%
Partisanship	Republican	26.8%	20.9%	16.5%	23.6%	16.5%
	Democrat	39.2%	33.8%	43.0%	44.7%	40.8%
	Independent	27.3%	36.8%	33.5%	22.1%	32.0%
	Other	2.1%	4.0%	4.0%	1.0%	2.9%
	No Preference	4.6%	4.5%	3.0%	8.5%	7.8%

Note: We report mean values (with SD in parentheses) for age and percentages for the other variables. Participants who selected the age category “60 or older” were assigned an age value of 60. Some category totals may not sum to 1 due to rounding.

Findings. We found a negative interaction between news summary valence and generation, *b* = -0.15, *SE* = 0.02, *t* = -6.29, *p* < 0.001, *95% CIs* [-0.20, -0.10]. In the “positive headlines” condition, comments became more negative across generations, *b* = -0.06, *SE* = 0.02, *t* = -3.59, *p* < 0.001, *95% CIs* [-0.10, -0.03]. However, the reverse was the case for participants in the "negative headlines" condition, *b* = 0.09, *SE* = 0.02, *t* = 4.91, *p* < 0.001, *95% CIs* [0.05, 0.12]. This generation x initial valence interaction did not significantly vary across the political articles and non-political articles, *b* = 0.08, *SE* = 0.05, *t* = 1.67, *p* = 0.10, *95% CIs* [-0.01, 0.17]. Valence changes across generations, along with detailed distributions, are illustrated in **Fig. S12**.

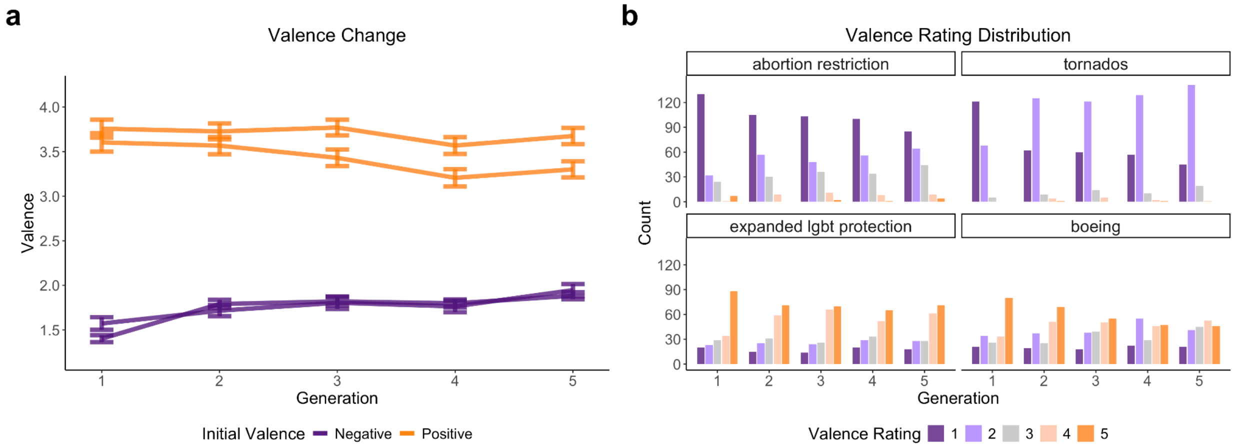

Fig. S12. Valence change across generations using VADER. a. Valence of participants' responses to different news headlines across generations. **b.** Valence distribution of participants' responses to different news headlines across generations, as rated by GPT-4.

3. Language Model Prompts

3.1. Language Model Prompts for Evaluating News Headlines

We used the GPT-4 API (version gpt-4o-2024-08-06) to evaluate news headlines with the following prompt:

> You're a helpful annotator. Consider this news headline: {HEADLINE}.\n Please rate this headline on positivity, impact, and partisanship.\n For positivity, use a -1 to 1 scale, where -1 is very negative and 1 is very positive. For impact, use a 0 to 1 scale, where 0 is very low impact and 1 is very high impact. For partisanship, use a -1 to 1 scale, where -1 is very partisan to the Democrats side and 1 is very partisan to the Republicans side.\n Return ONLY the ratings in the format: '0.5, 0.8, -0.3', which correspond to positivity, impact, and partisanship, respectively. Always return three ratings separated by commas. If you cannot determine a rating, return 'NA' for that rating. For example, '0.5, NA, -0.3'.

The placeholder {HEADLINE} was replaced with the actual news headline when calling the API.

3.2. Language Model Prompts for Evaluating Experiment Responses

Prompt for the main experiment. We used the GPT-4 API (version gpt-4-turbo-2024-04-09) to clean and screen responses with the following prompt:

> Determine if the provided text makes logical sense and looks like a social media comment in reaction to hearing this piece of news: {HEADLINE}. Note that such a social media comment may represent a personal reflection or argument that is triggered by this news, even though it doesn't directly mention the news. Here is the text: {TEXT}. If the text does not make logical sense and/or does not look like a social media comment reacting to the news, output 'NA'. If the text makes logical sense and looks like a social media comment reacting to the news, correct any typos and return the corrected version in English.

The placeholder {HEADLINE} was replaced with the actual news headline, and {TEXT} was replaced with the participant’s response when calling the API.

Prompt for the supplementary experiment. We used the GPT-4 API (version gpt-4o-2024-08-06) to clean and screen responses with the following prompt:

> Assume you are an assistant in a psychology study. The following is a social media post written by participants we would like you to encode. The content is their reaction to different news events. Please evaluate the positivity of this text on a scale from 1 to 5 (1 = very negative, 3 = neutral, 5 = very positive). Please note that you should only encode the degree of positivity directly expressed in the text and not infer it based on the event. You do not need to provide any justification. Please organize your response in the following format: Positivity: SCORE. SCORE should be a number between 1 and 5. Here is the text: {CLEANED_TEXT}

The placeholder {CLEANED_TEXT} was replaced with the participant’s response when calling the API.

Note that the prompt used in the supplementary experiment is slightly different from the one used in the main experiment, as the data cleaning processes for the two experiments were handled by different authors, who agreed to use different prompts as a way to increase robustness.

Supplementary References

1. Waller I, Anderson A. Quantifying social organization and political polarization in online platforms. Nature. 2021;600(7888):264–8.

2. Rajadesingan A, Budak C, Resnick P. Political discussion is abundant in non-political subreddits (and less toxic). In: Proceedings of the International AAAI Conference on Web and Social Media. 2021. p. 525–36.

3. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media. 2014. p. 216–25.

4. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

5. Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature human behaviour*, *7*(5), 812-822.

6. Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, *14*(2), 259-269.

7. Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. Trends in Cognitive Sciences, 27(10), 947-960.

8. Alves H, Koch A, Unkelbach C. Why good is more alike than bad: Processing implications. Trends Cogn Sci. 2017;21(2):69–79.

9. Koch A, Alves H, Krüger T, Unkelbach C. A general valence asymmetry in similarity: Good is more alike than bad. J Exp Psychol Learn Mem Cogn. 2016;42(8):1171.

10.Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

11. Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529-544.