Information Research - Vol. 31 No. iConf (2026)

Who gets blamed? Negativity bias in social media diffusion on Weibo

Tiantian Liang, Xi Wang, Mei Zhang, and Jian Tang

DOI: https://doi.org/10.47989/ir31iConf64147

Abstract

Introduction. This study examines negativity bias in social media diffusion by analysing Weibo discussions on women's marriage and fertility, focusing on how opinion valence shapes diffusion dynamics.

Method. We collected Weibo posts from January to March 2023, yielding 6,405 original posts and 97,090 diffusion chains after pre-processing. A large language model was used to code opinion polarity of original posts and first-level comments, and regional participant composition was measured to capture contextual heterogeneity.

Analysis. Diffusion depth was operationalised as chain length. Multilevel linear mixed-effects models were employed to estimate the effects of original opinion valence, with first-level comment polarity and regional dominance specified as moderators.

Results. Negative posts triggered deeper, longer diffusion while positive posts curtailed discussion spread. The stance of the first comment played a critical moderating role, that positive originals achieved wider diffusion when initial responses were negative. Regional context further moderated these effects, with eastern-user-dominated chains showing weaker amplification of negativity but stronger diffusion of positive content.

Conclusion. Negativity bias strongly shapes social media diffusion, but its effects are conditional on early audience responses and regional participant composition.

Introduction

For more than three decades, fertility in China was tightly bound by the one-child policy introduced in 1979 and enforced through the birth-permit system of the 1980s (J. Li, 1995). In 2016, the one-child rule was abolished, the permit requirement was formally scrapped in favour of simple birth registration, and couples were encouraged to have two children. By 2021, the ceiling was lifted again to three. Yet this policy reversal, from strict control to active encouragement, has not reversed demographic decline. Marriage rates continue to fall, and births remain at historic lows (H. Li et al., 2024). In response, public conversations about marriage and fertility have increasingly moved online, where negative framings and anxious narratives often dominate attention and shape wider discussion.

Within this online discourse, women's marriage and fertility choices stand at the center of contention. Labels such as '*leftover woman*' routinely trigger heated exchanges (Xu, 2021), and policy proposals debated during China's annual '*Two Sessions*'—the National People's Congress and the Chinese People's Political Consultative Conference—from 2022 to 2024 further intensified these discussions by challenging entrenched social norms, including marital requirements for birth registration and reproductive options for unmarried women. Yet the dynamics of these debates are rarely balanced. Attention-driven platforms disproportionately reward emotionally charged framings, which allows alarmist or stigmatising narratives to eclipse more constructive voices. In this environment, negativity bias becomes a defining mechanism that negative content captures greater visibility and spreads more widely (Bebbington et al., 2017), reinforcing prejudice and amplifying public anxiety around marriage and fertility.

A growing body of research has examined user behaviour on social media, ranging from participation motives and churn prediction to information diffusion and algorithmic amplification (Stieglitz & Dang-Xuan, 2013; X. Wang et al., 2025; Yu et al., 2024). Within this literature, negativity bias has consistently emerged as a powerful driver of engagement, with negative content more likely to attract attention, be shared, and shape perceptions (Rozin & Royzman, 2001; Takano et al., 2023). Studies of Weibo and other platforms have documented how emotionally charged discourse can fuel polarisation and distort public understanding (Serrano-Puche, 2021). However, existing scholarship leaves two important gaps. First, research has rarely situated negativity bias in the sensitive and policy-relevant context of marriage and fertility debates in China. Second, few studies have systematically modeled diffusion at the chain level to identify how negativity is moderated by early audience reactions or by the regional composition of participants. Addressing these limitations is crucial for understanding not only the general dynamics of online communication but also the particular challenges posed by demographic governance in contemporary China.

This study investigates negativity bias in the context of Weibo discussions on women's marriage and fertility. We analyse user comments and the diffusion chains they form to examine how negative opinions extend the reach of online debates. Particular attention is given to two moderators that have received limited empirical attention: the evaluative stance of the first comment and the regional composition of participants. The analysis shows that negativity bias generally lengthens diffusion chains, but its impact is conditional. Positive interventions at the outset of a chain can significantly dampen negative cascades, while chains dominated by users from economically developed regions, such as eastern regions, display weaker tendencies to amplify negativity.

The findings make a twofold contribution. Theoretically, the study extends research on negativity bias by identifying boundary conditions that limit its influence in online discourse. Practically, it provides insights for governance by highlighting how early interventions and regional context can shape the trajectory of sensitive debates, offering guidance for managing social media platforms and communicating demographic policy more effectively.

Literature review

Negativity bias

Negativity bias describes the asymmetry in human information processing whereby negative events, stimuli, or evaluations exert a stronger impact than neutral or positive ones (Vaish et al., 2008). Classic psychological research demonstrates that individuals are more attentive to negative cues, encode them more deeply in memory, and weigh them more heavily in judgment and decision-making (Rozin & Royzman, 2001). This disproportionate sensitivity manifests across a range of cognitive and affective processes, including attention allocation, emotional reactivity, and evaluative reasoning, leading to the well-documented principle that *'bad is stronger than good'* (Baumeister et al., 2001).

Social and behavioral research also confirms that negative experiences shape interpersonal relations and institutional trust more powerfully than positive ones (Baumeister et al., 2001; Fiske, 1980; Rozin & Royzman, 2001). For instance, one critical interaction with a peer or an institution can erode confidence disproportionately relative to multiple positive encounters (Rozin & Royzman, 2001). This asymmetry extends beyond individual cognition to collective behavior. For example, in collective decision-making, groups are more likely to focus on risks, conflicts, and failures than on opportunities or cooperative gains (Janis, 1982; Kahneman & Tversky, 1984).

Negativity bias in social media diffusion

Social media provides a communication environment in which negativity bias is not only reproduced but systematically amplified. Unlike traditional media, where editorial gatekeeping constrains information flows, social platforms such as Weibo and Twitter allow users to generate and disseminate content in real time at minimal cost. This combination of immediacy and low participation barriers enables emotionally charged content, particularly negatively framed posts, to gain rapid visibility and engagement, often before corrective information or more balanced perspectives can emerge (Vosoughi et al., 2018). Consequently, negative narratives tend to dominate trending topics, disproportionately shaping collective attention and public discourse.

The design of social media platforms further magnifies these tendencies. In the attention economy, algorithms prioritise content that generates high engagement, irrespective of its informational quality or social value (Liang, 2022). Negative content — being concise, vivid, and emotionally arousing — tends to elicit stronger reactions, such as outrage, ridicule, or moral condemnation, which translate into clicks, shares, and comments. Algorithmic curation then reinforces this cycle by pushing such content to wider audiences, producing a feedback loop that privileges negativity (GausenAnna et al., 2022). In this sense, social media does not merely reflect negativity bias but actively institutionalises it through its design and business logic.

At the same time, the social dynamics of online interaction provide fertile ground for the spread of negative attitudes. On Weibo and other platforms, users form participatory networks where opinions and emotions propagate through posting, commenting, and reposting. Conformity effects make individuals sensitive to the tone of peers in their networks, often leading them to adopt similar stances. As negative comments accumulate, groups may experience polarisation, shifting from moderate discussion to entrenched hostility against specific positions or populations. Prior research has identified that the first comment is particularly consequential, as the earliest response visible to all, it can anchor subsequent interpretations of the original post, set the emotional tone, and decisively influence whether the discussion evolves into constructive debate or a negative spiral (Garimella et al., 2017; S. Y. Lee & Kim, 2023; Macy et al., 2019).

Marriage and fertility discourse

Marriage and fertility have become highly contested topics in contemporary Chinese social media. Discussions of delayed marriage, declining fertility intentions, and the perceived burdens of childrearing are not only frequent but also emotionally charged. Existing research suggests that debates surrounding marriage and fertility are deeply intertwined with broader socioeconomic transformations, including housing affordability, gender inequality, and career pressures (Caucutt et al., 2002; He et al., 2024; W. Li, 2024). Social media platforms thus provide a salient arena in which these grievances and anxieties are articulated, circulated, and reshaped through everyday interactions.

Empirical studies further indicate that negative framings of marriage and fertility—such as '*marriage as a trap*', '*lying flat*' in response to family pressures, or the high costs of raising children—tend to circulate more widely than positive narratives (Ning et al., 2022). Such negative portrayals attract disproportionate attention and exert stronger influence on online opinion climates. As a result, public discourse increasingly privileges narratives of dissatisfaction and resistance, potentially reinforcing skepticism toward state-promoted ideals of marriage and pro-natalism.

From a diffusion perspective, these patterns can be partly explained by the dynamics of social contagion and algorithmic amplification. Exposure to critical or pessimistic commentary often elicits similar expressions among peers, generating cascades of disillusionment or resistance. At the same time, recommendation algorithms on major social media platforms systematically boost highly engaging content, accelerating the visibility of negative framings while crowding out more moderate voices (Husza´r et al., 2022; Milli et al., 2023). Consequently, the digital public sphere surrounding marriage and fertility is characterised less by balanced deliberation than by recurrent amplification of polarised and pessimistic narratives.

Research gaps

Although existing research has highlighted the salience of negativity bias and its amplification through social media diffusion, several important research gaps remain. First, most studies focus on Western contexts, leaving a limited understanding of how negativity manifests in non-Western settings, particularly within Chinese social media ecosystems where state regulation, platform governance, and cultural expectations shape discourse differently.

Second, the boundary conditions of negativity bias are still insufficiently understood. Prior work has rarely examined when and where negativity bias is likely to intensify or weaken. Social-economic contexts may create such boundaries, yet empirical evidence remains limited. This gap motivates a closer investigation into how socioeconomic disparities condition the diffusion of negativity in Chinese social media.

Hypotheses development and research model

Hypotheses development

Prior research consistently finds that negative content garners disproportionate attention and engagement on social media, spreading more rapidly and widely than neutral or positive content (Bebbington et al., 2017; Vosoughi et al., 2018). This '*negativity advantage*' is especially pronounced in emotionally sensitive domains such as marriage and fertility, where stigmatising framings, depictions of high personal cost, or expressions of grievance are more likely to resonate with audiences and trigger extended exchanges. Based on this reasoning, we hypothesize:

> **H1**: Negative original posts will be associated with greater diffusion depth compared to neutral or positive posts.

Group interactions on social media are highly susceptible to conformity effects, which can drive discussions toward polarisation. Users' attitudes and behaviors are easily influenced by others in their networks, often aligning with increasingly extreme positions. On Weibo, the attitude expressed in first-level comments is particularly consequential. As the earliest visible responses to an original post, these comments often set the tone for subsequent interactions and influence whether a discussion develops into constructive deliberation or devolves into a negative spiral (T. Wang et al., 2016). For instance, when the first comment directly contradicts the original post, it may provoke debate and accelerate polarisation. As negative responses accumulate, conversations can become more extreme, sometimes escalating into hostility toward specific groups. These dynamics underscore the guiding role of early comments in shaping both the depth of diffusion and the emotional orientation of the discussion. Based on this reasoning, we hypothesize:

> **H2**: Negative first-level comments moderate the relationship between original Weibo opinions and diffusion depth.

Beyond interpersonal dynamics, diffusion is also shaped by broader structural contexts. Homophily clusters users into relatively homogeneous groups, narrowing exposure and reinforcing locally shared norms (McPherson et al., 2001). Since regional categories often proxy underlying socio-economic differences—including levels of development, social inclusiveness, and digital literacy—the regional composition of participants within a diffusion chain may condition how negativity bias manifests in online discussions. Systematic differences across regions in social norms, value orientations, and discursive climates may create distinct structural conditions under which emotionally charged content is amplified or attenuated (Ottaviano & Peri, 2006; Roscoe et al., 2020; X. Wang et al., 2021). Thus, while our model operationalises this factor as regional dominance, it conceptually captures the socio-economic context in which online discourse evolves. Accordingly, we hypothesize:

> **H3**: The regional dominance of participants within a diffusion chain moderates the relationship between the original post's opinion valence and diffusion depth.

In summary, this study conceptualises negativity bias in original Weibo posts as the independent variable and diffusion depth as the dependent variable. Two moderators are incorporated: first-level comment opinions (H2) and regional characteristics of participants (H3). Diffusion depth is operationalised as the length of diffusion chains. Since a single post can generate multiple chains that share the same source, the data are inherently nested rather than independent. To address this, we adopt a multilevel linear mixed-effects model. The overall research framework is presented in fig. 1.

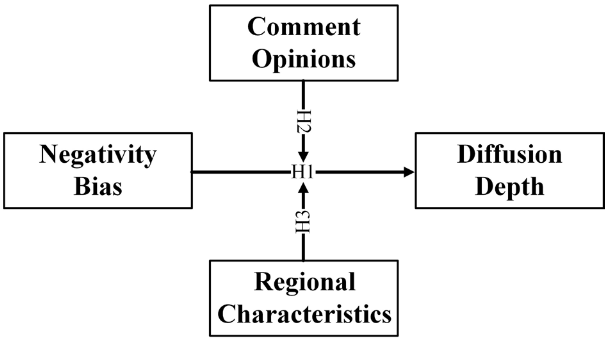

Figure 1. Research framework.

Model specification

We estimate diffusion depth at the chain level j, nested within original posts i, using multilevel linear mixed-effects models with a random intercept for each post. The baseline specification is:

$$\begin{aligned} \text{depth}_{ij} = \beta_0 + \beta_1\, originalOpinion_i \\ + \beta_2 Z_j + \beta_3(originalOpinioni \times Z_j) \\ + \gamma^{\top}\mathbf{controls}_{ij} + v_i + \varepsilon_{ij} \end{aligned} \tag{1}$$

where v_i captures unobserved post-level heterogeneity and ε_{ij} denotes the idiosyncratic error term.

Model 1 (H1). We set Z_j and the interaction term to zero. The coefficient β_1 tests whether the valence of the original post is associated with diffusion depth.

Model 2 (H2). We define $Z_j = followOpinion_j$, representing the sentiment of the earliest first-level comment in chain j. The interaction coefficient β_3 examines whether early audience responses moderate the relationship between the original post's opinion and diffusion depth.

Model 3 (H3). We define $Z_j = dominantRegion_j$, capturing regional dominance within the diffusion chain. In this specification, β_3 tests whether regional composition moderates the association between opinion valence and diffusion depth.

The dependent variable $depth_{ij}$ measures diffusion depth as the length of the diffusion chain. $originalOpinion_i$ and $followOpinion_j$ denote the opinion categories of the original post and the earliest first-level comment, respectively. Both variables are coded as categorical indicators (0 = negative, 1 = neutral, 2 = positive), with neutral serving as the reference category.

$dominantRegion_j$ captures the regional composition of participants within diffusion chain j, inferred from IP-based user location data. Drawing on China's well-established East–West developmental gradient, we operationalise regional context as Eastern dominance, which proxies for higher socioeconomic development, digital literacy, and social tolerance. For each chain, we compute the proportion of users from Eastern regions: the variable is coded as 3 (Eastern-dominant) if more than 50% of participants are from Eastern regions, 2 (parity) if Eastern and non-Eastern users are equally represented, and 1 (non-Easterndominant) otherwise. Non-Eastern dominance serves as the reference category.

$\mathbf{controls}_{ij}$ includes chain-level topical controls, measured by the average relevance of all comments to predefined thematic categories. Finally, v_i denotes the random intercept at the original-post level.

Research methods

Data

Weibo is one of the most influential social media platforms in China, enabling users to share text, images, and videos and to interact through reposts, likes, and comments. Due to its central role in disseminating trending topics and everyday communication, Weibo constitutes a valuable source for studying public opinion.

For this study, we collected data from January 1 to March 31, 2023, corresponding to the first quarter of 2023, focusing on the highly salient societal debate over women's marriage and fertility. We retrieved original posts containing the keywords '*unmarried women*', '*non-marital childbirth*', '*single motherhood*', and '*leftover women*', along with all associated comments. These thematic categories also serve as topics in calculating chain-level topical controls. It is important to note that a single original post may generate multiple diffusion chains, as illustrated in Figure 2.

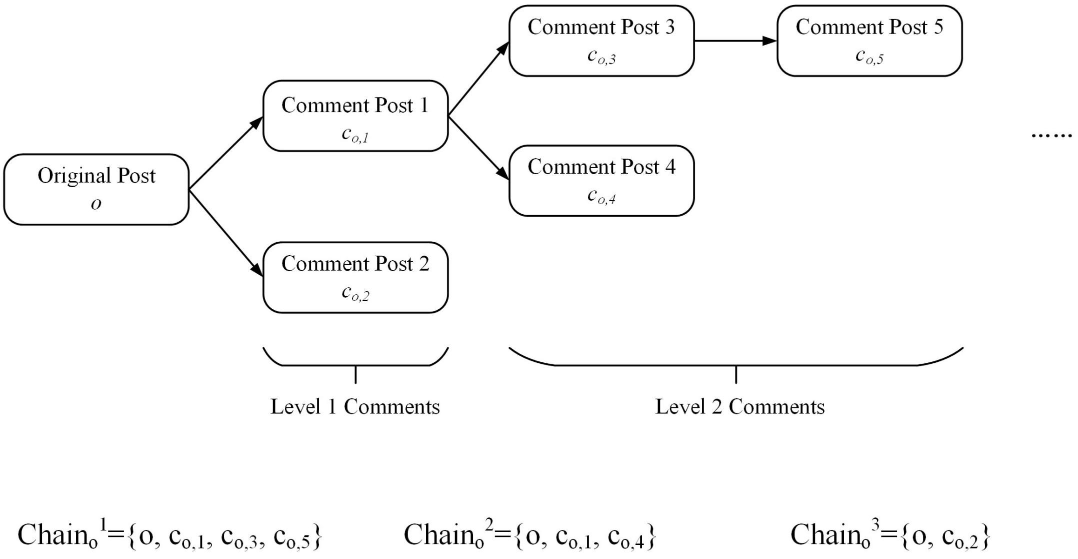

Figure 2. Illustration of a Weibo diffusion chain.

The dataset comprises 90,468 users and 142,245 records, including 18,729 original Weibo posts and 123,516 comments. Records from the eastern regions account for the majority, with 99,046 entries representing approximately 70% of the total. During preprocessing, duplicate records and original posts without diffusion chains (i.e., posts receiving no comments) were removed. After cleaning, the dataset retained 97,090 diffusion chains derived from 6,405 original posts. A detailed summary of the data is provided in Table 1.

Opinion Category	Original Post	Comment	Original Post(%)	Comment(%)
Negative	26,869	45,088	27.68	46.45
Neutral	7,849	26,262	8.08	27.05
Positive	62,367	25,724	64.24	26.50
Total	97,085	97,074	100.00	100.00

Table 1. Descriptive statistics.

Opinion classification by LLMs

To examine the diffusion of negative information, it is essential to establish clear opinion labels for both original posts and comments. In this study, we employed the GPT-4o-mini model provided by OpenAI to classify 129,561 processed Weibo records (original posts and comments), assigning each entry a single opinion label. To minimise value-laden bias, we define the classification framework ex ante. Content emphasising respect for individual freedom of choice is coded as *positive*, encompassing both the endorsement of traditional marriage and fertility norms, as well as support for remaining unmarried or childless. By contrast, content involving discrimination, stigmatisation, stereotyping, or the incitement of gender conflict is coded as *negative*.

Compared with manual annotation, which is costly and limited in scale, LLM-based classification allows large-scale, consistent, and context-sensitive coding of opinions. To implement this, we applied prompt engineering, a strategy that guides models through carefully designed input prompts. Common methods include zero-shot, few-shot, roleplaying, and chain-of-thought prompting (Kojima et al., 2022; Kong et al., 2024; Wei et al., 2022; Ye & Durrett, 2022). Zero-shot prompting, the most widely used, specifies the task in natural language without additional training, while few-shot prompting provides examples to shape outputs. Role-playing prompting simulates particular identities or roles, and chain-of-thought prompting elicits step-by-step reasoning to improve interpretability.

In multiple rounds of experimentation, this study combined few-shot, role-playing, and chain-of-thought prompting, iteratively refining the prompts, and ultimately established a relatively stable prompt design for the LLMs as follows:

> As an artificial intelligence system with expertise in language and sentiment analysis, your task is to analyse the following Chinese texts containing subjective descriptions. The texts are sourced from the Chinese social media platform Weibo and involve topics related to '*unmarried women*,' '*non-marital childbirth*,' '*older unmarried women*,' and '*single-parent childbirth*,' including both original posts and comments.

Step 1: Relevance Identification
Confirm whether the received text is related to the mentioned topics.

Step 2: Opinion Classification
If the text is relevant, further analyse its opinion and classify it into one of three categories: positive, neutral, or negative.

1. **Positive**: Includes statements that support gender equality, respect individual choices, or oppose gender discrimination. Specifically:
   - Expressions of respect for those who marry, give birth, remain unmarried, remain childless, or give birth as single parents; • Emphasis on the freedom and rights of personal choice (e.g., timing and partner choice for marriage/childbirth);
   - Challenges to gender stereotypes.
2. **Negative**: Includes statements that demean or deny gender equality, reinforce stereotypes, or hold negative attitudes toward unmarried individuals or single-parent births. Specifically:
   - Derogatory descriptions of unmarried or childless individuals;
   - Deliberate incitement of gender conflict, such as using extreme language to attract attention or expressing hostility toward men under the guise of '*feminism*,' or exhibiting ethnic discrimination.
3. **Special instructions**: • Texts containing sarcasm or irony but fundamentally supporting gender equality should be classified as positive; • Texts that appear positive on the surface but substantively reinforce gender inequality should be classified as negative.

Step 3: Output requirements
Please output the following four items:

- relevance: Topic relevance classification (0 = unrelated, 1 = neutral, 2 = related);
- opinion: Opinion category classification (0 = negative, 1 = neutral, 2 = positive);

- positive prob: Probability that the text expresses a positive opinion, ranging from [0,1];
- negative prob: Probability that the text expresses a negative opinion, ranging from [0,1].

> Ensure that the sum of positive prob and negative prob equals 1, and report the probabilities with four decimal places of precision.

Using this prompt design, the model demonstrated improved accuracy in distinguishing stances within complex contexts. For example, in preliminary experiments without prompt optimisation, the statement '*The most important thing for women is to find a good partner and have a happy family*' was incorrectly classified as positive. After optimisation, however, the model identified the underlying gender stereotype and correctly categorised it as negative, consistent with the classification framework adopted in this study.

Following the finalisation of the prompt scheme, large-scale classification and annotation were conducted across the dataset. As summarised in Table 1, negative stances accounted for nearly half of all comments (46.45%), substantially higher than their proportion in original posts (27.68%). This pattern suggests that negative expressions are more readily elicited in interactive contexts, providing preliminary evidence for the propagation of negativity bias in social media discussions.

To assess the reliability of the LLM-based opinion classification, we randomly sampled 500 unique original posts for manual annotation and compared the human labels with the LLM outputs. We report both percentage agreement and Cohen's kappa (κ), which adjusts for chance agreement and is widely used for categorical coding (Cohen, 1960). The LLM achieved an agreement rate of 89.98%. Cohen's kappa was $\kappa = 0.8247$ with a 95% confidence interval of [0.7585, 0.8909] ($p < 0.001$), indicating near-perfect agreement under common interpretive guidelines (Cohen, 1960).

Results

Baseline effects of original post valence (H1)

Model 1 in Table 2 tests the effect of original post valence on diffusion depth. Results indicate that negative originals generate significantly longer diffusion chains than neutral ones ($\beta = 0.0994, p < 0.05$), whereas positive originals do not differ from the baseline. These findings support Hypothesis H1 and substantiate a negativity bias, suggesting that negative content is more likely to attract attention and stimulate interaction, thereby extending diffusion.

Variables	Model 1	Model 2	Model 3
originalOpinion: Negative	0.0994** (0.0462)	0.0512 (0.0566)	0.1243* (0.0706)
originalOpinion: Positive	0.0229 (0.0425)	-0.0314 (0.0514)	-0.2469*** (0.0645)
followOpinion: Negative		0.4631*** (0.0493)	
followOpinion: Positive		0.8272*** (0.0578)	
dominantRegion: Equal Regions			-0.6524*** (0.0579)
dominantRegion: Eastern Dominance			0.1680*** (0.0611)
originalOpinion × *followOpinion*: Negative × Negative		0.0674 (0.0562)	
originalOpinion × *followOpinion*: Negative × Positive		-0.0627 (0.0656)	
originalOpinion × *followOpinion*: Positive × Negative		0.1318** (0.0523)	
originalOpinion × *followOpinion*: Positive × Positive		-0.0001 (0.0604)	
originalOpinion × *dominantRegion*: Negative × Equal Regions			0.0619 (0.0674)
originalOpinion × *dominantRegion*: Negative × Eastern Dominance			-0.1181* (0.0708)
originalOpinion × *dominantRegion*: Positive × Equal Regions			0.2829*** (0.0617)
originalOpinion × *dominantRegion*: Positive × Eastern Dominance			0.3753*** (0.0649)
N	**97046**	**97046**	**97046**

Table 2. Comparison of results across three models.

Notes: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Moderation by early comment (H2)

Model 2 in Table 2 incorporates the valence of the first comment as a moderator. Relative to a neutral first comment, both negative ($\beta = 0.4631$, $p < 0.01$) and positive ($\beta = 0.8272$, $p < 0.01$) first comments are associated with significantly longer diffusion chains. Crucially, the interaction between a positive original post and a negative first comment is positive and significant ($\beta = 0.1318$, $p < 0.05$). This indicates that a negative first reply neutralises and reverses the slight dampening otherwise linked to positive originals, yielding a net effect of approximately +0.1004 (−0.0314 + 0.1318). Other interactions are not statistically significant. These results support Hypothesis H2.

In substantive terms, whether a *positive* original spreads depends on how the conversation begins. When the first reply is neutral, positive originals do not lengthen diffusion (the coefficient is small and not significant). By contrast, when the first reply is *negative*, the significant interaction indicates that the thread moves into deeper diffusion on the net. This pattern is consistent with a negativity-sensitive audience response, that an early negative reply to a positive stance appears to raise contestation and invite participation, thereby extending the chain. Notably, we do not find reliable evidence that the effect of *negative* originals varies with first-comment valence, suggesting that the moderation concentrates on cases where a positive stance meets an early negative reply.

Moderation by regional composition (H3)

Model 3 in Table 2 introduces regional dominance as a moderator. In the main effects, negative originals are positively associated with diffusion depth ($\beta = 0.1243$, $p < 0.10$), while positive originals are significantly shorter in diffusion than neutral ones ($\beta = -0.2469$, $p < 0.01$), further confirming the baseline negativity advantage observed in Model 1.

Turning to the moderation analysis, the baseline category is chains dominated by nonEastern users. Three interactions reach statistical significance: negative originals in Eastern-dominant chains ($\beta = -0.1181$, $p < 0.10$), positive originals in equal-region chains ($\beta = 0.2829$, $p < 0.01$), and positive originals in Eastern-dominant chains ($\beta = 0.3753$, $p < 0.01$). Substantively, this means that the extension effect of negative originals observed elsewhere is essentially cancelled in Eastern-dominant chains (net effect ≈ 0.006). Conversely, positive originals, which otherwise shorten diffusion, flip to significantly lengthening chains when either equal or Eastern regional composition prevails (net effects of ≈ +0.036 and ≈ +0.128, respectively). These results support Hypothesis H3.

In substantive terms, the regional context conditions how negativity bias unfolds. In Eastern-dominant chains, the greater inclusiveness and cultural diversity of participants reduce the tendency for negative posts to sustain extended diffusion, while simultaneously encouraging engagement with positive originals. By contrast, in non-Eastern chains, the baseline negativity bias remains evident. Therefore, diffusion depth is shaped not only by post valence and early comments but also by the socio-economic composition of the audience, revealing boundary conditions to the otherwise robust amplification of negativity.

These moderation effects underscore the role of Eastern regions, where higher economic development and social inclusiveness foster more tolerant attitudes toward marriage and fertility debates. Public responses in these contexts are more supportive of positive views and less prone to repeatedly discuss negative content. Importantly, economic prosperity not only underpins inclusiveness but also shapes online participation patterns, reinforcing the divergence in diffusion dynamics between Eastern and non-Eastern contexts.

Robustness check

Because diffusion depth is a count variable, we further conducted robustness checks using generalised linear mixed-effects models (GLMMs) with a log link (Y. Lee & Nelder, 1996). The results are reported in Table 3 (Models 4–6). Across all specifications, negative original posts generate significantly deeper diffusion chains than neutral ones, whereas positive originals exhibit weaker or negative associations with diffusion depth.

Variables	Model 4	Model 5	Model 6
originalOpinion: Negative	0.0353** (0.0157)	0.0189 (0.0203)	0.0784*** (0.0236)
originalOpinion: Positive	0.0043 (0.0152)	-0.0178 (0.0190)	-0.0679*** (0.0224)
followOpinion: Negative		0.1727*** (0.0493)	
followOpinion: Positive		0.3004*** (0.0209)	
dominantRegion: Equal Regions			-0.2593*** (0.0209)
dominantRegion: Eastern Dominance			0.0629*** (0.0209)
originalOpinion × *followOpinion*: Negative × Negative		0.0216 (0.0208)	
originalOpinion × *followOpinion*: Negative × Positive		-0.0259 (0.0237)	
originalOpinion × *followOpinion*: Positive × Negative		0.0464** (0.0194)	
originalOpinion × *followOpinion*: Positive × Positive		0.0022 (0.0218)	
originalOpinion × *dominantRegion*: Negative × Equal Regions			0.0244 (0.0244)
originalOpinion × *dominantRegion*: Negative × Eastern Dominance			-0.0579** (0.0241)
originalOpinion × *dominantRegion*: Positive × Equal Regions			0.1017*** (0.0224)
originalOpinion × *dominantRegion*: Positive × Eastern Dominance			0.1161*** (0.0222)
N	**97046**	**97046**	**97046**

Table 3. Comparison of GLMM results across three models.

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The moderating role of the first comment is likewise stable. A negative first response to a positive original consistently amplifies the diffusion depth, while other interaction terms remain weak or statistically insignificant. In addition, the moderation effect by regional composition persists under the GLMM framework. Specifically, in diffusion chains dominated by users from Eastern regions, the amplification effect of negative originals is significantly attenuated. Conversely, positive originals reverse their baseline disadvantage and significantly extend diffusion in both Eastern-dominant and regionally balanced chains.

Taken together, these robustness checks indicate that the observed negativity bias and its boundary conditions are not driven by distributional assumptions of the dependent variable, thereby strengthening the robustness and credibility of the empirical findings.

Conclusion and future research

By analysing Weibo discussions on women's marriage and fertility, this study demonstrates how negativity bias shapes diffusion chains and how its effects are moderated by user attitudes and regional contexts. Negative originals significantly lengthen diffusion chains, while positive originals either shorten or fail to extend them. The influence of negativity is stronger than that of positivity, indicating that negative content draws more user attention and stimulates interaction, whereas positive sentiment dampens subsequent commenting.

The findings also reveal that regional context plays a critical role. In Eastern-dominant chains, negative posts lose their advantage, while positive posts extend diffusion, reversing the baseline pattern. This suggests that in economically developed and socially inclusive regions, user engagement is more responsive to positive sentiment, highlighting boundary conditions for the otherwise robust amplification of negativity.

These results carry important implications for platform governance. Because negative content tends to diffuse widely, unchecked amplification may distort online discourse and public perceptions of sensitive issues. Strengthening monitoring mechanisms while safeguarding freedom of expression is therefore essential. At the same time, the greater inclusiveness observed in Eastern regions can be leveraged by elevating positive content and expanding access to constructive voices, thereby reducing the dominance of negativity and fostering healthier, more balanced dialogue within social media ecosystems.

Despite these contributions, several limitations should be acknowledged. First, regional dominance, used as a proxy for socioeconomic context, cannot fully capture within-region heterogeneity, such as urban–rural divides or population mobility. Second, diffusion dynamics may also be influenced by chain-level factors, including posting time, user network size, and algorithmic recommendation mechanisms, which are not systematically controlled due to data constraints. Finally, the analysis is confined to Weibo and the first quarter of 2023, which may limit the generalisability of the findings. Future research could address these limitations by incorporating finer-grained contextual measures, cross-platform data, and longer observation windows to further assess the robustness and boundary conditions of negativity bias.

Acknowledgments

We sincerely thank the reviewers for their constructive feedback and insightful sugges- tions. This research was supported by the Beijing Social Science Foundation Project (Grant No. 24GLC056).

About the Authors

Tiantian Liang is a graduate student majoring in Management Information Systems at the School of Information, Central University of Finance and Economics, China. Her re- search interests include social media analytics. She can be contacted at 2020312168@email.cufe.edu.cn

Xi Wang is a Professor at the School of Information, Central University of Finance and Economics, China. Her research focuses on user behavior analysis, business intelligence, and intelligent information systems. She can be contacted at xiwang@cufe.edu.cn

Mei Zhang is a Professor at the School of Sociology and Psychology, Central University of Finance and Economics, China. Her research focuses on online psychology and digital behavior. She can be contacted at zhangmeisd@163.com

Jian Tang is an Associate Professor at the School of Information Management, Nanjing University, China. His research interests include information behavior, human-computer interaction, and open collaboration. He can be contacted at jiantang@nju.edu.cn

References

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. Review of General Psychology.

Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: Evidence of a negativity bias in the social transmission of information. Evolution and Human Behavior, 38(1), 92–101. https://doi.org/10.1016/j.evolhumbehav.2016. 07.004

Caucutt, E. M., Guner, N., & Knowles, J. (2002). Why Do Women Wait? Matching, Wage Inequality, and the Incentives for Fertility Delay. Review of Economic Dynamics, 5(4), 815–855. https://doi.org/10.1006/redy.2002.0190

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. Journal of Personality and Social Psychology, 38(6), 889– 906. https://doi.org/10.1037/0022-3514.38.6.889

Garimella, K., De Francisc iMorales, G., Gionis, A., & Mathioudakis, M. (2017). Mary, Mary, Quite Contrary: Exposing Twitter Users to Contrarian News. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, 201–205. https://doi.org/10.1145/3041021.3054737

GausenAnna, LukWayne, & GuoCe. (2022). Using agent-based modelling to evaluate the impact of algorithmic curation on social media. ACM Journal of Data and Information Quality. https://doi.org/10.1145/3546915

He, Y., Abdul Wahab, N. E. T., & Muhamad, H. (2024). Factors impacting fertility anxiety among Chinese young women with marital status differences. Heliyon, 10(1), e23715. https://doi.org/10.1016/j.heliyon.2023.e23715

Husza´r, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences, 119(1), e2025334119. https://doi.org/10.1073/pnas.2025334119

Janis, I. L. (L. (1982). Groupthink : Psychological studies of policy decisions and fiascoes. Boston : Houghton Mifflin.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. American Psychologist, 39(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Proceedings of the 36th International Conference on Neural Information Processing Systems.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024, June). Better zero-shot reasoning with role-play prompting. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers) (pp. 4099–4113). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.228

Lee, S. Y., & Kim, J.-H. (2023). What makes people more polarised? The effects of anonymity, being with like-minded others, and the moderating role of need for approval. Telematics and Informatics, 76, 101922. https://doi.org/10.1016/j.tele. 2022.101922

Lee, Y., & Nelder, J. A. (1996). Hierarchical Generalised Linear Models [Publisher: [Royal Statistical Society, Oxford University Press]]. Journal of the Royal Statistical Society. Series B (Methodological), 58(4), 619–678. https://www.jstor.org/stable/ 2346105

Li, H., Tang, J., & Qiao, J. (2024). China's declining fertility rate. BMJ, 385, q1000. https://doi.org/10.1136/bmj.q1000

Li, J. (1995). China's one-child policy: How and how well has it worked? A case study of hebei province, 1979-88. Population and Development Review, 21(3), 563–585. https://doi.org/10.2307/2137750

Li, W. (2024). Do surging house prices discourage fertility? Global evidence, 1870–2012. Labour Economics, 90, 102572. https://doi.org/10.1016/j.labeco.2024.102572

Liang, M. (2022). The end of social media? How data attraction model in the algorithmic media reshapes the attention economy. Media, Culture & Society. https://doi.org/10.1177/01634437221077168

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarisation. Science Advances, 5(8), eaax0754. https://doi.org/ 10.1126/sciadv.aax0754

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology, 27(1), 415–444. https://doi.org/ 10.1146/annurev.soc.27.1.415

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2023). Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media [Version Number: 6]. https://doi.org/10.48550/ARXIV.2305.16941

Ning, C., Wu, J., Ye, Y., Yang, N., Pei, H., & Gao, H. (2022). How Media Use Influences the Fertility Intentions Among Chinese Women of Reproductive Age: A Perspective of Social Trust. Frontiers in Public Health, 10, 882009. https://doi.org/10. 3389/fpubh.2022.882009

Ottaviano, G. I., & Peri, G. (2006). The economic value of cultural diversity: Evidence from US cities. Journal of Economic Geography, 6(1), 9–44. https://doi.org/10. 1093/jeg/lbi002

Roscoe, R. D., Chiou, E. K., & Wooldridge, A. R. (Eds.). (2020). Advancing diversity, inclusion, and social justice through human systems engineering. CRC Press, Taylor & Francis Group.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Sage Journals. https://doi.org/10.1207/S15327957PSPR0504 2

Serrano-Puche, J. (2021). Digital disinformation and emotions: Exploring the social risks of affective polarisation. International Review of Sociology.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems, 29(4), 217–248. https://doi.org/10.2753/MIS0742-1222290408

Takano, M., Nakazato, K., & Taka, F. (2023). Dynamics of discrimination and prejudice via two types of social contagion [Publisher: Elsevier]. Applied Mathematics and Computation, 448, 127916. https://doi.org/10.1016/j.amc.2023.127916

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development [Place: US Publisher: American

Psychological Association]. Psychological Bulletin, 134(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online [Publisher: American Association for the Advancement of Science TLDR: A large-scale analysis of tweets reveals that false rumors spread further and faster than the truth, and false news was more novel than true news, which suggests that people were more likely to share novel information.]. Science. https://doi.org/10.1126/ science.aap9559

Wang, T., Chen, Y., Wang, Y., Wang, B., Wang, G., Li, X., Zheng, H., & Zhao, B. Y. (2016). The power of comments: Fostering social interactions in microblog networks. Frontiers of Computer Science, 10(5), 889–907. https://doi.org/10.1007/ s11704-016-5198-y

Wang, X., Zhang, Y., Wang, S., & Zhao, K. (2021). Migrant Inflows and Online Expressions of Regional Prejudice in China. Public Opinion Quarterly, 85(1), 123–146. https://doi.org/10.1093/poq/nfab004

Wang, X., Zhang, Y., Li, H., & Zuo, Z. (2025). Treat or quit: Churn prediction in online health communities based on inverse reinforcement learning. Electronic Commerce Research. https://doi.org/10.1007/s10660-025-10000-8

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems.

Xu, Y. (2021). Understanding the phenomenon of leftover women in China. 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021), 2205–2209. https://doi.org/10.2991/assehr.k.211220.380

Ye, X., & Durrett, G. (2022). The unreliability of explanations in few-shot prompting for textual reasoning. Proceedings of the 36th International Conference on Neural Information Processing Systems.

Yu, W.-J., Hung, S.-Y., Yu, A. P.-I., & Hung, Y.-L. (2024). Understanding consumers' continuance intention of social shopping and social media participation: The perspective of friends on social media [Publisher: North-Holland]. Information & Management, 61(4), 103808. https://doi.org/10.1016/j.im.2023.103808