Content Generation on Social Media: The Role of Negative Peer Feedback[*]

Varad Deolankar[†] Jessica Fong[‡] S. Sriram[§]

July 26, 2023

Abstract

Social media platforms rely on their users for the supply of content, which in turn, is important for driving engagement from other users. In recent years, user-generated content (UGC) on social media platforms has received increased scrutiny for creating a polarizing environment wherein users post extreme opinions and/or receive negative backlash from dissenters. In this paper, we ask how receiving negative peer feedback, in the form of downvotes, affects user-generated content (UGC). We focus on two aspects of UGC: (a) willingness to post (i.e., incidence) and (b) the extremity of subsequent posts (i.e., intensity). Using data from Reddit, we find that receiving negative (and positive) peer feedback increases a user's subsequent posting activity, relative to no feedback. However, unlike positive feedback, receiving negative peer feedback moderates extreme sentiments such that when the initial opinion is extreme, users reduce the intensity of their subsequent opinions upon receiving negative feedback. These effects of negative feedback are consistent with users attempting to maintain their reputation. Our findings highlight the potential benefits of negative feedback in promoting UGC creation. They also suggest that negative peer feedback can serve as the whip that regulates the polarization of opinions by encouraging users to moderate their tone.

Keywords— Social Media, Polarization, Customer Retention, Platform Design

[*] We are thankful for comments from Justin Huang, Megan Hunter, Rupali Kaul, Kosuke Uetake, and from seminar participants at Yale SOM, University of Michigan, and the Marketing Science 2023 conference. The usual disclaimer applies.

[†] University of Michigan, varadd@umich.edu

[‡] University of Michigan, jyfong@umich.edu

[§] University of Michigan, ssrira@umich.edu

1 Introduction

Social media platforms have become ubiquitous in modern society, with 4.6 billion users worldwide as of 2022.[1] Such platforms provide a space for people to connect, share opinions and information, and discuss a wide variety of topics. Social media platforms typically promote this process by incorporating peer feedback. For example, Facebook users can "like" posts, and Reddit users can give peers "badges". Existing research has shown that positive peer feedback, which can be expressed as upvotes (Jin et al., 2015), badges (Burtch et al., 2022; Gallus, 2017; Huang et al., 2022), comments (Wang and Majeed, 2022), and the number of followers (Toubia and Stephen, 2013), can increase content generation and promote user engagement.

However, social media platforms often also solicit negative feedback. For example, Reddit users can downvote posts, and YouTube users can dislike videos. Although negative feedback can be useful for moderating content and improving recommendation algorithms, it can also impact the production of user-generated content (UGC). On one hand, negative feedback has the potential to diminish intrinsic motivation (Deci and Cascio, 1972), thereby potentially reducing UGC output. On the other hand, negative feedback can generate attention (Berger et al., 2010) and thus motivate users to increase UGC production. In addition to impacting the volume of UGC, negative feedback may also shape the nature of content itself. For instance, if negative feedback is undesirable, users may adapt their future content to align more closely with popular opinion in order to avoid criticism. This can lead to the suppression of unpopular content or opinions. Alternatively, negative feedback can cause users to become defensive (Nussbaum and Dweck, 2008), or induce users to strengthen their (unpopular) opinions because it can generate more attention (Cheng et al., 2014). These effects on content can have implications for echo chamber formation and polarization on social media platforms.

In this paper, we study the role of negative feedback in driving the creation of UGC and ask the following questions. First, how does receiving negative feedback affect the likelihood that the user creates content in the future, relative to receiving positive feedback or no feedback? Second, how does negative feedback impact the content, conditional on posting again? More specifically, does negative feedback moderate or magnify the strength of the opinions expressed (i.e., intensity) in the subsequent content? Answers to these questions will enable us to comment on the role of negative feedback on echo chamber formation and polarization.

We explore these questions in the context of Reddit. Reddit is an ideal setting to study these

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, accessed May 2023.

questions for several reasons. First, users' posts accumulate feedback in the form of both upvotes and downvotes, which are aggregated to produce a "score" for each post. Furthermore, users receive feedback via "karma", which is a function of the scores of all their posts combined. Karma and posts' scores are visible to both the content creator and to other Reddit users. Changes in score and karma serve as objective measures of feedback, with a reduction in either indicating negative feedback. Although feedback can be received in other ways (i.e., text replies), upvotes and downvotes provide a straightforward metric to quantify the valence of feedback. For this reason, we use changes in users' karma and scores received for their comments as measures of feedback. Second, discussions on Reddit tend to be opinion-based. This subjectivity of UGC is a noteworthy feature of the context as it allows considerable latitude for users to adjust their content in response to feedback from their peers. As a result, we can infer how feedback influences the nature of content produced and its implications for the polarization of opinions.

In order to perform our empirical analysis, we collected data on the posting behavior, the score of each post and their karma every day for a sample of Reddit users. This enables us to track the content creation of each user as well as the feedback they receive from their peers (i.e., the change in post-level scores and user-level karma) over time. We focus on commenting behavior because comments make up the vast majority of UGC on Reddit.[2]

Measuring the causal effect of feedback on subsequent posting behavior can be challenging because of the following reasons. First, users may differ in terms of their intrinsic propensity to create content and in their tendency to discuss controversial topics or deliberately post provocative content (e.g., "trolls"). Such users are also likely to receive more (negative) feedback. Second, users tend to be more active on social media on certain days (e.g., weekends). We are likely to to see more UGC and feedback on such days. We can control for these two confounds that are common within an individual over time and common across individuals within a day using user and day fixed effects, respectively. Third, unobserved factors, such as political events or the release of a new movie, may impact both the amount and intensity of feedback generated on certain topics. Such unobserved factors that vary across topics and over time can pose a serious challenge to inferring a causal relationship between feedback and subsequent UGC.

To overcome the latter challenge, we use an identification strategy that is akin to a regression discontinuity approach. In particular, we assume that users are likely to feel the effect of feedback more strongly when the score or karma crosses certain thresholds, such a left digit change (Bizer and Schindler, 2005; Anderson and Simester, 2003; Strulov-Shlain, 2022; Thomas and Morwitz, 2005;

[2] The other form of UGC is submissions, which is the parent post, such as a question, that users can comment on. Comments account for 95% of UGC in our data.

List et al., 2023; Husain et al., 2021; Lacetera et al., 2012), or a sign change.[3] For example, because of left digit bias, negative feedback is more likely to be salient if the user's karma decreases from 101 to 99, as opposed to a decrease from 102 to 100, despite both receiving the same amount of negative feedback (-2). Similarly, negative feedback is more salient if a post's score decreases from 2 to -1 than from 4 to 1. Thus, we assume that after controlling for the level of feedback, karma, and score, the changes in the left digit of karma or the sign of a comment's score are exogenous.

We find that negative feedback increases content creation. More specifically, for a given magnitude of negative feedback, users are more likely to create a comment when the negative feedback is felt more saliently (i.e., the left digit in their karma decreases). We find a similar effect for positive feedback; users are more likely to create a comment when positive feedback is felt more saliently (i.e., the left digit in their karma increases). Therefore, both positive and negative feedback positively impact the probability of posting again. This is consistent with findings in other settings—that engagement from peers, regardless of the valence, can increase the focal user's engagement with the platform (Cheng et al., 2014; Eckles et al., 2016; Mummalaneni et al., 2022; Zhu et al., 2022). In addition, the increase in posting activity in response to negative feedback is not driven by users defending their original comment. Rather, we find that users are more likely to comment under other submissions in "subreddits", which are individual communities within Reddit, where they had not contributed to recently.

In addition to increasing incidence, we find that negative feedback can moderate the intensity of the content. We used Google's Natural Language API to quantify the intensity of the text. We observe that if a user receives negative feedback on a comment expressing intense sentiment toward a particular topic, they tend to moderate the intensity of their tone in future posts about that topic, provided they post about it again. More precisely, subsequent mentions of the topic are less intense if the original comment's score decreases and changes sign, compared to the user's other comments that decrease in score but do not change signs. We do not find any evidence of moderation effects from positive feedback.

Our paper contributes to the literature on social media platforms—in particular, feedback systems and their effect on UGC creation. Our results support the notion that negative feedback may have additional benefits beyond improving recommendation algorithms. Specifically, we find that negative feedback increases the propensity of users to create content, especially in communities they had not participated in recently. In addition, negative feedback moderates extreme intensity, potentially promoting more civil discussions. These findings are encouraging for social media platforms

[3] We verify that this assumption is consistent in our setting using placebo tests, as described in Sections 4.1.3 and 4.2.5.

that are considering whether to introduce tools that enable users to provide negative feedback. Our paper also contributes to the ongoing discourse on echo chambers and polarization in social media. Existing discussions predominantly revolve around the role of recommendation algorithms in echo chamber formation and the subsequent amplification of polarization (some examples include Sunstein (2009); Pariser (2011); Dandekar et al. (2013); Flaxman et al. (2016); Levy (2021); Cinelli et al. (2021); Rafieian and Yoganarasimhan (2023)).[4] Our research introduces the idea that peer feedback systems can serve as an additional factor influencing polarization. Our results indicate that such systems have the potential to mitigate polarization by moderating the intensity of user interactions.

The remainder of the paper is structured as follows. Section 2 describes the related literature. In Section 3, we provide background on our empirical setting and present summary statistics. Section 4 presents our empirical analysis for both incidence and intensity. Section 5 discusses the potential underlying mechanisms and implications for platforms, echo chambers and polarization. We conclude with the limitations and directions for future research in Section 6.

2 Related Literature

2.1 Feedback and UGC Creation

Our paper relates to the literature that studies UGC creation on social media platforms. Extant literature has identified several reasons why users create content, such as social network ties (Peng et al., 2018; Shriver et al., 2013), intrinsic and image-related utility (Zhang and Zhu, 2006; Moe and Schweidel, 2012; Toubia and Stephen, 2013). Our paper most closely relates to the stream of research that focuses on how feedback impacts UGC.

The literature has primarily focused on positive feedback, which is often received in the form of platform and peer recognition. This has been studied in various contexts, such as Wikipedia (Gallus, 2017), online education (Denny, 2013), an image-sharing social network (Huang and Narayanan, 2020), and Reddit (Burtch et al., 2022; Lu et al., 2022). These studies have consistently documented a positive causal relationship between recognition and user-generated content (UGC) production. However, the impact of recognition on nature of the content itself has yielded mixed findings. Burtch et al. (2022) found that platform recognition tends to encourage exploitation, leading to new UGC that closely resembles the previously recognized content. In contrast, Huang et al. (2022) document that peer recognition fosters exploration, resulting in subsequent content that differs from the recognized content.

[4] One notable exception is Yoganarasimhan and Yakovetskaya (2022), which explores how individuals' choices in sharing news articles contribute to polarization on social media.

Another stream of literature has examined the impact of peer feedback more broadly. Using a peer encouragement design, Eckles et al. (2016) demonstrate that receiving feedback from peers, such as “Likes” and comments, increases a user’s propensity to create posts and engage with other users’ content on Facebook. Mummalaneni et al. (2022) investigate the heterogeneous effects of peer engagement on Twitter. Their findings reveal that while feedback, including “Likes,” replies, and “Retweets”, does not significantly influence most Twitter users, there exists a small subset of highly responsive users for whom feedback has a notable impact. These responsive users increase time spent on Twitter, generate more new posts, and send a greater number of “Favorites” to other users. However, these studies do not differentiate between positive and negative feedback, which is one of the primary contributions of our research. Furthermore, we expand upon this literature by examining how feedback shapes the subsequent content produced in response to such feedback.

The literature on the impact of negative feedback on user-generated content (UGC) is relatively limited, and the existing findings regarding the effect on subsequent content quality are inconclusive. *Prima facie*, it might seem as though the effect of negative peer feedback would just be the opposite of corresponding positive feedback. Therefore, if positive peer feedback increases production of UGC, negative feedback should decrease it. Using observational data from a restaurant review platform, Zhu et al. (2022) find that both positive and negative feedback from peers increase a user’s likelihood of creating future content. Furthermore, negative feedback is associated with longer reviews that incorporate more pictures, which in turn receive more positive feedback. Cheng et al. (2014) examine the effect of negative feedback on commenting behavior in news websites (e.g., *CNN.com*). Through matching posts with similar textual qualities and users’ post histories, they find that users who receive negative feedback contribute more. However, contrary to Zhu et al. (2022), they observe that subsequent posts are of lower quality, as measured by human-labeled text quality scores.

Our paper contributes to this nascent literature by examining the impact of negative feedback in the context of a social media platform. Additionally, we explore the effect of negative feedback on the intensity of subsequent content. This aspect is particularly significant given the ongoing discussions surrounding echo chamber formation and polarization on social media platforms. As discussed below, we conjecture that negative feedback could potentially shape the formation of echo chambers on social media platforms and influence the degree of polarization.

2.2 Echo Chambers and Polarization

Previous research has identified several human tendencies that contribute to polarization and echo chambers on social media: homophily, which refers to the inclination to interact primarily with similar individuals (Kandel, 1978; McPherson et al., 2001), and confirmation bias, which involves seeking and favoring information that aligns with pre-existing beliefs (Freedman and Sears, 1965; Frey, 1986; Stroud, 2008). These tendencies, amplified by algorithmic recommendations, can lead to polarization and echo chambers on social media (Sunstein, 2009; Pariser, 2011). Recent studies have explored the role of recommendation systems in social media polarization. For example, Dandekar et al. (2013) show that an algorithm that recommends the item that is most relevant to a user on the basis of a PageRank-like score (Page et al., 1999), is always polarizing. Levy (2021) study the consumption of news on social media and document that social media algorithms limit exposure to counter-attitudinal news and thus increase polarization.

Our paper expands the existing literature by introducing the concept that peer feedback systems can play a significant role in influencing content polarization. We conjecture that negative feedback can potentially have two opposing effects on the propensity of users to create subsequent content. On the one hand, negative feedback can deter users from contributing to the space where they received the negative feedback. Under such a scenario, negative feedback could potentially trigger the exit of users with contrarian points of view, thereby contributing to the formation of echo chambers. On the other hand, negative feedback (like positive feedback) might be viewed as a form of attention and encourage users to create content.

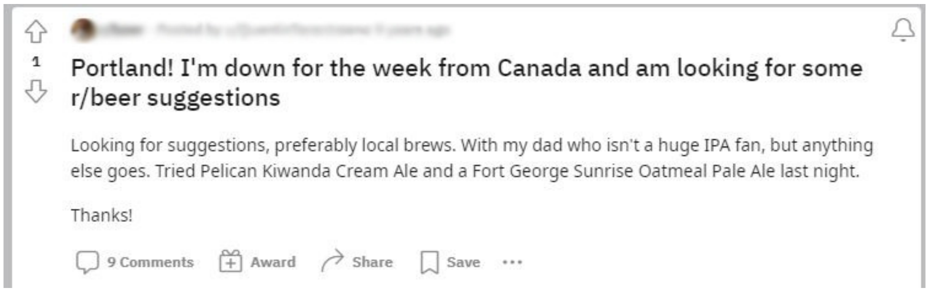
We conjecture that a similar tension might exist when considering the effect of negative feedback on the tone of subsequent content. While negative feedback might encourage users to moderate their tone, it could also induce them to become defensive (Nussbaum and Dweck, 2008), potentially taking more extreme positions. In light of this ambiguity, we empirically investigate how negative peer feedback on prior content affects the intensity of subsequent content, shedding light on the potential role of peer feedback systems as drivers of content polarization. As a result, our findings offer valuable insights into leveraging peer feedback systems to mitigate content polarization on social media.

3 Setting and Data

The empirical context of our paper is Reddit, a social media platform with a forum-style discussion structure. Users contribute to Reddit by posting content in communities called “subreddits”. Content can be either a “submission”, wherein a user starts a new topic for discussion, or a comment that users can make under a submission. Figure 1 displays an example of a “submission”, and Figure 2 displays examples of comments under a submission. In the remainder of our paper, we refer to a “post” as either a submission or a comment.

Reddit is an ideal setting for our study for the following reasons. First, often referred to as “the

Figure 1: Example of A Submission



[redacted]

1

Portland! I'm down for the week from Canada and am looking for some r/beer suggestions

Looking for suggestions, preferably local brews. With my dad who isn't a huge IPA fan, but anything else goes. Tried Pelican Kiwanda Cream Ale and a Fort George Sunrise Oatmeal Pale Ale last night.

Thanks!

front page of the Internet", Reddit is one of the most popular websites in the world, with over 50 million active users per day, and over 13 billion comments and posts.[5] The volume of UGC allows us to explore the effect of feedback across a variety of topics. In addition, the topics discussed on Reddit are often opinion-based (e.g., what is your favorite restaurant in Portland?). This provides a setting where users have the latitude to take multiple perspectives without being "wrong". That is, Reddit is a setting in which many different opinions about the same topic can exist. Second, Reddit allows users to provide positive *and* negative peer feedback on UGC, a feature crucial to our study. Users can upvote or downvote content, and the platform summarizes these at two different levels of aggregation. At the post level, Reddit reports the "score", which is the net difference between the number of upvotes and downvotes. At the user level, the platform provides an overall measure of reputation in the form of submission and comment "karma" points, which, in turn, are functions of the total feedback received by the user over their lifetime. A user's submission (comment) karma increases with the upvotes she receives on her submissions (comments) and decreases with the downvotes on her submissions (comments).[6] The score of each post is publicly displayed next to or under the post as shown in Figures 1 and 2.

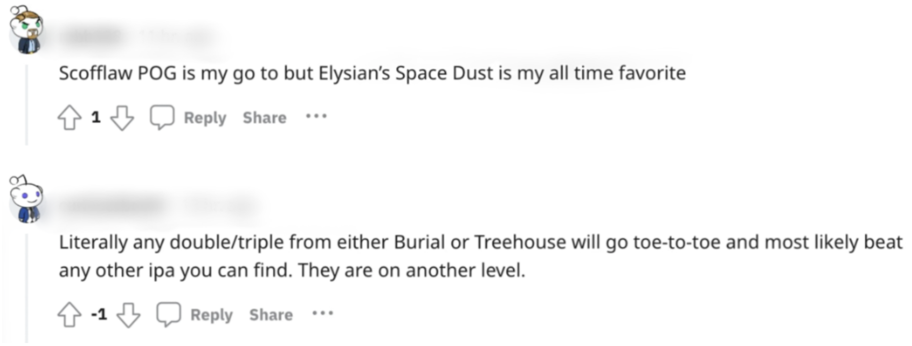
On Reddit, a post can receive peer feedback in a multiple ways, including obtaining text responses either through comments and getting upvotes or downvotes.[7] We focus on upvotes and downvotes because they provide an objective measure of feedback.

[5] See https://www.redditinc.com/ and https://www.statista.com/statistics/443332/reddit-monthly-visitors, accessed December 2022.

[6] A user's comment karma is bounded below by -100. The exact formula for determining karma from upvotes and downvotes is not revealed by Reddit. More information on how karma is calculated can be found here: `https://reddit.zendesk.com/hc/en-us/articles/204511829-What-is-karma-`, accessed May 2023.

[7] Posts can also receive "badges", but there is no negative equivalent. Users can also receive private messages, which we do not observe.

Figure 2: Examples of Comments



[redacted]

Scofflaw POG is my go to but Elysian's Space Dust is my all time favorite

[redacted]

Literally any double/triple from either Burial or Treehouse will go toe-to-toe and most likely beat any other ipa you can find. They are on another level.

3.1 Data Collection

We collected our sample by first randomly selecting 20 popular subreddits.[8] We then selected the 100 most recent submission creators in those subreddits and collected their posting activity on a daily basis over a period of 41 days—from May 22, 2022 to July 3, 2022.[9] This provided us with a sample of 1,400 users.[10]

For each user, we observe the following time-invariant attributes: their account creation date, and the number of comments they had created, their submission karma and comment karma prior to the data collection period. During the data collection period, we observe their submission and comment karma at the daily level. We also observe each post the user creates during our data collection period. For each post, we observe the author, the type of post (submission or comment), the content, and the subreddit in which the content was created. Importantly, we also observe each post's score on each day, from the date it is posted until the end of our data collection period. This enables us to track how the score for each post changes over time, which is our metric for feedback (i.e., a decrease in score means the user received negative feedback). Note that Reddit displays a "fuzzed" score for each post, which is the true score (i.e., difference between upvotes and downvotes) plus noise.[11] Like the users, we observe only the fuzzed score, which is the actual score plus some random noise, and not the number of upvotes or downvotes. The random noise is redrawn each time

[8] "Popular" subreddits are those provided in the subreddit "ListOfSubreddits",`https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits/`, accessed December 2021.

[9] Data on users, their submissions, and their comments were collected using the Python Reddit API Wrapper (PRAW).

[10] This sample is smaller than 2,000 because not all subreddits had 100 unique submission creators and not all users posted during the observation period.

[11] Reddit fuzzes scores to prevent score manipulation by bots. More information can be found at `https://www.reddit.com/r/help/wiki/faq/#wiki_why_do_the_number_of_votes_change_when_i_reload_a_page.3F`, accessed May 2023.

the webpage loads, which implies that two users may see different fuzzed scores for the same post at the same time. Therefore, the score that we observe may be different from the score that the user observes. That is, our observed scores may include measurement error. We discuss the implications of score "fuzzing" for our identification strategy in Section 4.2.2. To our understanding, a user's karma is *not* fuzzed.

3.2 Sample Descriptives

Our final sample contains 140,110 posts in 6,247 unique subreddits from the 1,400 users. In Table 1, we present the statistics of our sample.

Table 1: Sample Summary Statistics

Statistic	Min	Q1	Median	Mean	Q3	Max
N submissions created per day	0	0	0.024	0.129	0.098	22.122
N comments created per day	0	0.146	0.561	2.527	2.122	57
Comment karma	-100	172	1163	13,446.7	6,332	2,497,871
Δ Comment karma	-752	2	7	50.080	28	7,574

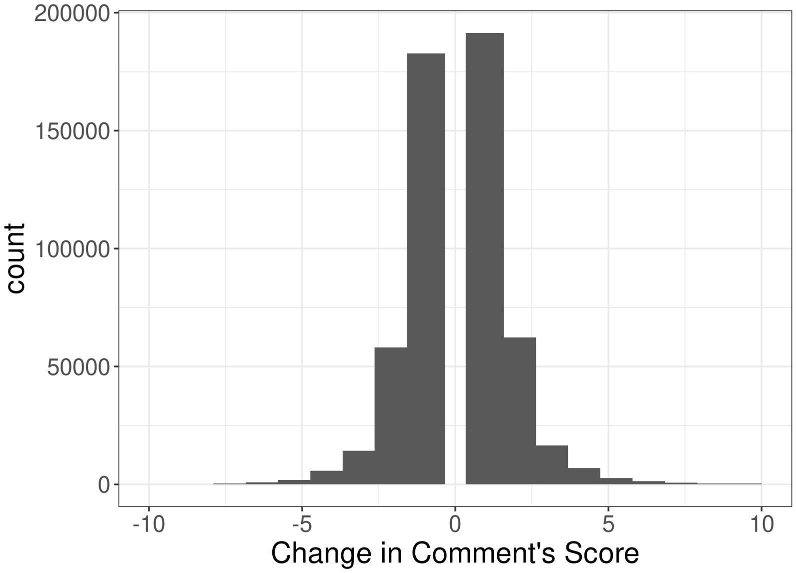
This table summarizes user-level characteristics. The first two rows report the number of submissions and comments created per user per day, respectively. The third row reports the distribution of users' comment karma on any given day. The last row reports the changes in their comment karma per day ($Karma_{it} - Karma_{i,t-1}$), conditional on any change in karma.

The first two rows of Table 1 show that the vast majority of posts in our sample are comments: on average, a user creates 0.129 submissions and 2.527 comments on a given day. In fact, 95% of the posts in our sample are comments. For this reason, we focus on comments for our analysis.

Rows 3 and 4 of Table 1 display users' comment karma and the changes in their comment karma. As shown in row 3, there is substantial heterogeneity across users in terms of their karma levels on any given day. Moreover, as row 4 indicates, on average, a user earns 50 comment karma each day conditional on earning any karma. This suggests that upon receiving feedback, on average, users receive positive feedback on their comments. However, there is also substantial heterogeneity in this dimension. Users also receive negative feedback in the form of a decrease in comment karma; the largest day-to-day decrease in comment karma in our sample is -752.

Next, we describe the fluctuation in comment scores generated by users within our sample throughout the observation period. These comments receive feedback in the form of daily changes in their scores. Figure 3 illustrates the distribution of these day-to-day score changes across all comments in our data. The distribution indicates that comments receive both positive and negative feedback, as evidenced by increases and decreases in their scores, respectively. In fact, among comments that receive feedback, 42% experience negative feedback on a given day.

Figure 3: Distribution of Feedback Received by Comments



This figure displays the distribution of the day-to-day change in score across all comments, conditional on there being a change in score.

4 Empirical Analysis

In this section, we describe how we measure the effects of negative feedback on subsequent UGC creation along two dimensions. In Section 4.1, we consider the effect of receiving negative feedback on a user's propensity to post (i.e., incidence). In Section 4.2, we investigate how receiving negative feedback impacts the intensity of subsequent comments posted by the user.

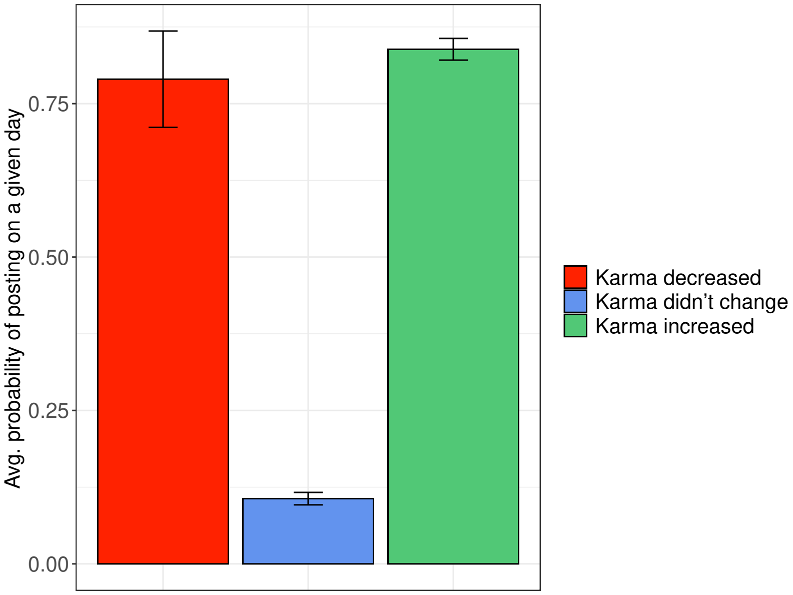
4.1 Effect of User Feedback on Incidence

Recall that users in our data create 2.527 comments per day, on average (see Table 1). Hence, a user is likely to receive feedback on multiple comments on a given day. Our objective is to establish a causal relationship between the feedback they receive across these comments and their subsequent posting behavior. As previously noted, Reddit provides an aggregate metric of a user's reputation in the form of their karma. These points fluctuate daily depending on the total number of upvotes and downvotes a user receives across all their comments. We measure feedback as the changes in comment karma during a given period of time. A reduction in comment karma signifies negative feedback, while an increase represents positive feedback.

4.1.1 Correlation between Feedback and Incidence

Before presenting the identification strategy, we first describe the correlation between changes in comment karma and subsequent posting activity. The average probability of a user creating a comment on any given day is 0.351. Figure 4 displays the average probability that a user creates a comment on a given day as a function of whether the user's comment karma has decreased, increased, or remained the same since the previous day. These data show that, on average, both negative and positive feedback are associated with an increased propensity to post, relative to receiving no feedback. At the same time, we do not observe a statistically significant difference in the propensities to post upon receiving positive or negative feedback.

Figure 4: Correlation between Feedback and the Probability of Posting



This figure displays the average probability that a user creates a comment on day t if the user's comment karma had decreased, increased, or didn't change between $t-1$ and t.

While these patterns may be suggestive of a causal relationship between receiving feedback and subsequent posting behavior, there are several other explanations. First, users may differ in terms of their time-invariant intrinsic propensity to post. Such heterogeneity might arise due to differences among users in terms of their level of motivation to create content on the platform. For example, heavy creators of UGC are likely to receive more peer feedback. We can control for this time-invariant, user-level heterogeneity with user fixed effects.

Second, the propensity of users to create content could vary over time. Such periods of high (or low) UGC creation may also coincide with the volume (and potentially the nature) of feedback that they receive. For example, users may be more likely to post on weekends but are also more likely to receive more peer feedback—and potentially more negative feedback—due to the higher level of user activity in general. We can control for such temporal trends that are shared by all users with time (day) level fixed effects.

Finally, even if we can control for cross-sectional and temporal heterogeneity, we may encounter confounding factors that vary over time within a user. We can control for the most obvious ones, such as whether the user posted in the previous day $t-1$ (e.g., users who post more recently are more likely to get feedback and may also be more likely to post again), karma levels (e.g., users with better reputations may be more likely to receive feedback and are more likely to post), and the absolute amount of feedback. We control for the absolute amount of feedback to account for the possibility that the amount of feedback can correlate with the topic of the comment and also the propensity to post again. Furthermore, as will become evident subsequently, this particular set of controls is important for our identification strategy.

We examine whether the correlations between feedback and incidence persist after controlling for cross-sectional and temporal heterogeneity, and the aforementioned variables that change within-user over time. To this end, we estimate the following logistic regression:

$$\begin{aligned}\mathbb{1}\{Comment_{it}\} = {} & \alpha_i + \alpha_t \\ & + \beta_1 \mathbb{1}\{NegativeFeedback_{it}\} + \gamma_1 \mathbb{1}\{PositiveFeedback_{it}\} + X'\Delta + \epsilon_{it}, \quad (1)\end{aligned}$$

where $\mathbb{1}\{Comment_{it}\}$ is an indicator for whether the user i posts a comment on day t. We include individual and calendar day fixed effects, α_i, and α_t. X denotes a matrix of controls, which include whether the user commented on the previous day ($\mathbb{1}\{Comment_{i,t-1}\}$), i's karma on the previous day ($Karma_{i,t-1}$, $Karma^2_{i,t-1}$), and the absolute value of amount of feedback ($|Feedback_{it}|$, $Feedback^2_{it}$) that user i receives between day $t-1$ and t.

We report the estimates of Equation (1) in Table 2. These results reinforce the earlier finding that receiving both positive and negative feedback are associated with a higher propensity to post (relative to receiving no feedback) in the subsequent period.

Table 2: Correlation between Feedback and Incidence

Dependent Variable: Model:	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.697^{***} (0.0963)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.837^{***} (0.0502)
Controls	✓
User fixed effects	✓
Date fixed effects	✓
Observations	49,197
Pseudo R^2	0.48997

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table displays the logistic regression estimates of Equation 1. Standard errors are clustered at the user level. Coefficient estimates for the controls are omitted for readability. Table C.1 in the Appendix reports all estimates.

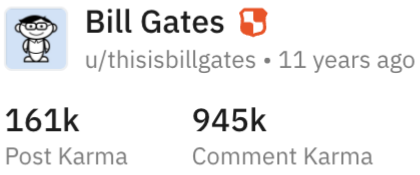
4.1.2 Identification Strategy

Despite controlling for individual and time level differences and user experience over time, there may still remain additional sources of omitted variable bias, which are in the form of unobserved shocks that vary within individuals over time. Such omitted variables can contaminate the causal effect estimates if they are correlated with the feedback that a user receives as well as their subsequent posting behavior. One such example is a change in interests. Over time, an individual may become more interested in politics and thus begin posting more about political topics. However, political posts are often more controversial than non-political posts and may lead to more negative feedback. Another example is topic-level shocks. Suppose an individual posts about a political topic (e.g., abortion) at time t, and a political event (e.g., the overturning of Roe v. Wade) occurs at $t + 1$. This event can impact the feedback the individual receives on his post made at time t, as well as his propensity to post after $t + 1$.

In order to address the concern about such omitted variables, similar to Fong and Hunter (2022), we adopt a regression discontinuity approach that exploits the idea that for a given level of feedback, changes in karma are likely to be more salient when karma crosses certain thresholds. Consider a scenario A in which a user's karma decreases from 104 to 100, and scenario B in which his karma decreases from 103 to 99. Despite receiving the same amount of feedback in both scenarios (-4), we posit that the negative feedback is likely to feel more salient to the user in scenario B than in scenario A because of left digit bias. Left digit bias is a well-documented phenomenon in which individuals tend to pay more attention to the leftmost digits and less attention to the rightmost digits when evaluating a number (Bizer and Schindler, 2005). This bias has been documented across a variety of contexts, such as retail prices (Anderson and Simester, 2003; Thomas and Morwitz, 2005; Strulov-Shlain, 2022), car mileage (Lacetera et al., 2012), and the quality of donor kidneys (Husain et al., 2021). Therefore, our identifying assumption is that controlling for the magnitude of feedback and karma, a left digit change in karma is uncorrelated with unobservables that impact posting incidence. Examples of unobserved attributes that correlate with posting incidence (and intensity) are visibility of the user's posts and the comments she receives, both of which may be correlated with the feedback the user receives. For example, one may expect that positive feedback (e.g., more upvotes) for a post may elicit more positive comments, as well as increase its visibility. However, after controlling for the amount of feedback, the occurrence of comments and changes in visibility are unlikely to be correlated with a change in the left digit of a user's karma. This assumption holds as long as a post does not elicit more comments or enhance visibility based on the left digit of the author's karma.

There are two assumptions behind this identification approach: (a) a left digit change indeed makes feedback more salient and (b) the user notices these changes. As regards (a), in addition to relying on the evidence documented elsewhere in the literature regarding left digit changes drawing more attention, we also perform a placebo test to verify this assumption. The results from this analysis, which we describe in Section 4.1.3 supports this idea. As regards (b), the way Reddit clearly displays the information about karma points (see Figure 5 for an example of how submission and comment karma are displayed) makes it more likely that users notice any changes therein. Nevertheless, we cannot guarantee that all users observe when their left digit changes. Therefore, we interpret these effects as an intent-to-treat (ITT) effect.

Figure 5: Example of How Karma is Displayed



Bill Gates

u/thisisbillgates • 11 years ago

161k
Post Karma

945k
Comment Karma

As seen from Reddit's desktop interface, a user's post (submission) karma and comment karma can be seen when hovering over a username.

Given this identification strategy, we estimate the following logistic regression:

$$\begin{aligned}\mathbb{1}\{Comment_{it}\} = \alpha_i + \alpha_t \\ + \beta_1 \mathbb{1}\{NegativeFeedback_{it}\} + \beta_2(\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}) \\ + \gamma_1 \mathbb{1}\{PositiveFeedback_{it}\} + \gamma_2(\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}) \\ + X_{it}'\Delta + \epsilon_{it}, \quad (2)\end{aligned}$$

$\mathbb{1}\{LDChanged_{it}\}$ is an indicator for whether the leftmost digit of user i's karma at day t is different than the leftmost digit of his karma at $t-1$. X_{it} denotes a matrix of controls, which include whether the user commented on the previous day ($\mathbb{1}\{Comment_{i,t-1}\}$); i's karma on the previous day ($Karma_{i,t-1}$, $Karma^2_{i,t-1}$); the amount of feedback ($Feedback_{it}$, $Feedback^2_{it}$), and the karma's distance from the nearest threshold where the left digit can change ($DistFromThreshold_{i,t-1}$, $DistFromThreshold^2_{i,t-1}$). Controlling for

$DistFromThreshold_{i,t-1}$ accounts for the possibility that users with karma closer to a threshold may post differently than those who are farther from a threshold. For example, a user with a karma of 199 ($DistFromThreshold_{i,t-1} = 1$) may post more or post content more likely to get positive feedback in an attempt to cross the threshold as compared to a user with a karma of 150 ($DistFromThreshold_{i,t-1} = 50$). We also include individual and date fixed effects, α_i and α_t.

The terms β_1 and γ_1 represent the change in a user's propensity to post when they receive negative and positive feedback in the previous period, respectively. However, as discussed above, we cannot interpret these as causal effects. Rather, the coefficients of interest are β_2 and γ_2, which correspond to the causal effects of negative and positive feedback, respectively, on subsequent content creation. An estimate of $\beta_2 > 0$ would indicate that users are more likely to comment if they received negative feedback *and* the left digit of their karma changes (i.e., the decrease in karma is more salient), relative to when they received negative feedback and the left digit of their karma did not change (i.e., the decrease in karma is less salient). Similarly, $\gamma_2 > 0$ indicates that positive feedback (i.e., a more salient *increase* in karma) increases posting propensity.

4.1.3 Results

We present the estimates of Equation (2) in Table 3. The estimates in column 1 show that the coefficient for the interaction between receiving negative feedback (i.e., a reduction in comment karma) and a left-digit change is positive and statistically significant. Thus, users are more likely to comment when negative feedback (i.e., a decrease in comment karma) is felt more saliently than in instances when it is less salient. The coefficient for the interaction between receiving positive feedback (i.e., an increase in comment karma) and a left-digit change is also positive and statistically significant. Thus, receiving either positive or negative feedback increases the production of content by users.

Table 3: Effect of Feedback on Incidence

Dependent Variable:	$\mathbb{1}\{Comment_{it}\}$		
Model:	(1)	(2)	(3)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.639^{***}	2.331^{***}	2.613^{***}
	(0.0990)	(0.1004)	(0.1529)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	1.169^{**}	0.9411^{*}	1.145^{**}
	(0.4941)	(0.5326)	(0.4992)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.805^{***}	2.228^{***}	2.811^{***}
	(0.0491)	(0.0503)	(0.0789)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.6512^{***}	0.4001^{**}	0.6923^{***}
	(0.1386)	(0.1624)	(0.2013)
User fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Controls	✓	✓	✓
Sample	All	$\lvert\text{Feedback}\rvert \leq 50$	DistFromThreshold ≤ 50
Observations	49,197	46,723	24,310
Pseudo R^2	0.49056	0.47023	0.41587

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table displays the logistic regression estimates of Equation 2. Standard errors are clustered at the user level. Coefficient estimates for the controls are omitted for readability. Table C.2 in the Appendix reports all estimates.

The premise behind our identification strategy is that there is something special about feedback that induces a left-digit change that makes it more salient than one that does not induce such a change. We verify this idea by conducting a placebo test wherein we assign individuals to the treatment of a left-digit change when their karma gets close (within 2 points) to a left-digit change but the left digit does not actually change. If a left digit change is especially salient, we should find a smaller effect in the placebo test. The estimates in Table 4 show that the coefficient of $\mathbb{1}\{LDChangedPlacebo\}$ is not statistically significant. This suggests that the feedback associated with a left digit change is indeed perceived differently by users, relative to when feedback decreases by a similar amount but the left digit does not change.

Next, we address three potential concerns with the estimates in Table 3 column 1. The first

Table 4: Effect of Feedback on Incidence: Placebo test

Dependent Variable: Model:	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.693^{***} (0.0976)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChangedPlacebo_{it}\}$	0.2648 (0.8369)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.834^{***} (0.0503)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChangedPlacebo_{it}\}$	0.3061 (0.3198)
User fixed effects	✓
Date fixed effects	✓
Controls	✓
Observations	49,197
Pseudo R^2	0.48999

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table displays the logistic regression estimates of Equation 2 using a placebo treatment. Standard errors are clustered at the user level. Coefficient estimates for the controls are omitted for readability. Table C.3 in the Appendix reports all estimates.

concern is that the average effect may be driven by outliers—users that receive a large amount of feedback, such as when their comment becomes viral. In such instances, karma is more likely to undergo a left digit change, and the user is also more likely to post in the current period. To address this concern, we verify that the association persists when we consider instances that do not experience a large magnitude of feedback. In particular, we restrict our analysis to instances where the magnitude of the feedback (both positive and negative) is less than 50. We present the results from this analysis in Column 2 of Table 3. The results reinforce our original results that both positive and negative feedback positively impact subsequent UGC creation.

The second concern is that users who are closer to the threshold where a left digit change might occur may behave differently than those who are farther away. Such a relationship would imply that the user's distance from the threshold could be correlated with them experiencing a left digit change and also their propensity to post. Although we control for the distance from the threshold in Equation (2), we can further address this concern by limiting the sample to only those who are closer to the threshold: observations where karma is within a distance of 50 from a left-digit change.[12] The results in column 3 of Table 3 suggest that the effects persist in this subsample.

The third concern is that users are more likely to experience a left-digit change when their karma

[12] 50 is the median distance to the threshold.

has fewer digits (i.e., two digits vs. three); the amount of feedback that is required to induce a left-digit change is greater when the karma has more digits. Although we control for the magnitude of feedback in our analysis, the parameterization might not fully capture the intricate relationship between the magnitude of the feedback and subsequent posting behavior. Therefore, we include fixed effects for the number of digits in $Karma_{i,t-1}$. We find that the effects, reported in Table 5, are robust.

Table 5: Effect of Feedback on Incidence: Fixed Effects for Number of Digits in Karma

	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.639^{***} (0.0990)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	1.190^{**} (0.4983)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.806^{***} (0.0491)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.6597^{***} (0.1405)
User fixed effects	✓
Date fixed effects	✓
Number of Digits fixed effects	✓
Controls	✓
Observations	49,197
Pseudo R^2	0.49065

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level. Coefficient estimates for the controls are omitted for readability. Table C.4 in the Appendix reports all estimates.

Effect of feedback on where users post

We also investigate the source of the increased propensity to post upon receiving negative feedback. Such sources include a change in the propensity to post in (a) the same submissions and communities in which they receive the negative feedback and/or (b) different communities. Of these, a decrease in (a) would suggest that the overall increase in a user's propensity to post upon receiving negative feedback was driven by an increase in (b). Under such a scenario, negative feedback might have deterred users from staying engaged in the same space (i.e., submission or the subreddit where their comment received negative feedback) and therefore exacerbate the formation of echo chambers. We test which pathway(s) contribute to the overall effect that we documented in Table 3 by considering

Table 6: Effect of Feedback on Incidence: Where Users Post

Dependent Variables: Model:	Samesubmission (1)	Samesubreddit (2)	Diffsubreddit (3)
$\mathbb{1}\{NegativeFeedback_{it}\}$	0.5680**	1.472***	2.296***
	(0.2388)	(0.1018)	(0.0887)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	-0.3198	-0.6384	1.473***
	(1.132)	(0.4908)	(0.4099)
$\mathbb{1}\{PositiveFeedback_{it}\}$	0.4493***	1.573***	2.414***
	(0.1085)	(0.0598)	(0.0507)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.5709*	0.0505	0.5539***
	(0.3064)	(0.1015)	(0.0946)
User fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Observations	17,709	34,004	49,687
Pseudo R^2	0.20665	0.25038	0.33359

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table displays the estimates of a logistic regression similar to Equation 2. The dependent variables are whether the user creates a comment in a different submission than the ones they commented on in the previous seven days, the user creates a comment in the same submission they commented on in the previous seven days, in a different subreddit than the one(s) they commented on in the past seven days, and in the same subreddit(s) but different submission(s) in the previous seven days, respectively. Standard errors are clustered at the user level. The number of observations vary across columns because of the fixed effects. All coefficient estimates are reported in Table C.5 in the Appendix.

subsequent commenting activities by users in (i) the same submission, (ii) the same subreddits but different submissions, and (iii) different subreddits. Of these, (i) and (ii) capture whether a user stays engaged in the same space (submission or subreddit, respectively) where they received negative feedback, while (iii) reflects their propensity to stay active on the platform but in a different space.

In order to implement this idea empirically, we need to link feedback to specific comments posted by a user. Since we use a change in karma (a user-level metric) to measure feedback, we need to link this change to specific comments. In our data, we observe that 99.8% of feedback is received within three days of comment creation. Therefore, if we consider the set of comments created by a user a few days prior to experiencing negative changes to their karma, we can isolate the corresponding content towards which this negative feedback was directed. Of course, if this set has multiple comments, we cannot uniquely identify the specific comment(s) that received negative feedback.

We present the results from this analysis in Table 6. As discussed above, the dependent variable in the first column is whether the user creates a comment in the same *submission* as the ones they commented on in the prior seven days.[13] In column 2, the dependent variable is whether the user comments in at least one of the *subreddits* in which they commented in the previous seven days. We do not find a statistically significant effect of negative feedback on a user's propensity to stay engaged in the same space. This implies that we do not find evidence of negative feedback exacerbating echo chambers. However, users are more likely to engage in different subreddits than the ones they commented in previously (column 3). Therefore, the average increase in commenting activity is due to engagement in different spaces than the ones in which they received the negative feedback. That is, negative feedback increases the users' engagement with the overall platform, rather than within specific communities. We find a similar effect for positive feedback.

4.2 Effect on Intensity of the Content

Next, we assess how receiving negative peer feedback changes the nature of an individual's subsequent content, conditional on posting. Because our sample contains a wide variety of topics, we cannot focus on commonly-used attributes of text, such as political leaning or direction (i.e., positive versus negative). Therefore, we focus on the intensity of content, which is applicable to many contexts. To illustrate, we consider the statement "Blue is a nice color" to be less intense than "Blue is the best color". We describe how we quantify a comment's intensity in Section 4.2.1.

While assessing whether and how receiving negative peer feedback changes the intensity of the content, we assume that the impact would be realized only for the topic that such feedback is directed

[13] We also conduct a robustness check that considers comments created in the previous three days. We find similar results. The results are not reported in the paper but can be provided upon request.

towards. For example, if a user receives negative feedback towards a comment that mentions topic A, we expect that the negative feedback may impact his language towards topic A in the future but not his language towards an unrelated topic B. This assumption implies we should conduct our analysis at the topic level. However, on Reddit, feedback is given for a comment, rather than for a specific topic within a comment; if a user receives downvotes for a comment that mentions topics A *and* B, we cannot determine whether the downvotes are directed towards the user's mention of A or his mention of B.

Therefore, we focus on how the feedback towards a focal *comment* impacts the intensity of a user's subsequent comments involving the topic(s) mentioned in the focal comment. As a result, in this section, we consider feedback at the comment level. We define the feedback that a comment receives on a given period t to be the change in the comment's score between t and $t-1$ (i.e., a comment receives negative feedback at t if the comment's score had decreased from $t-1$ to t).

4.2.1 Measuring Intensity

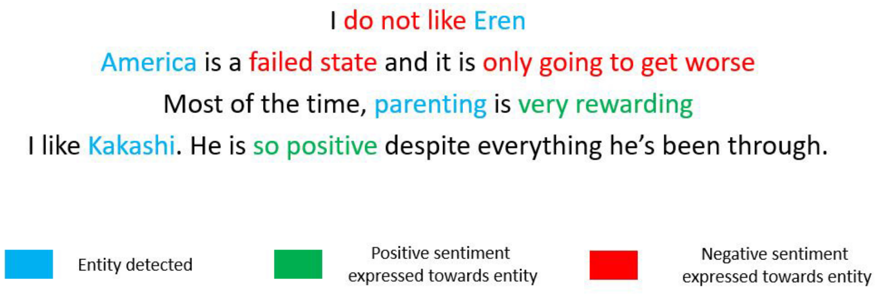
We use Google's Natural Language Processing (NLP) Application Programming Interface (API) to measure the intensity of text. In particular, Google's NLP API uses aspect-based sentiment analysis (also known as entity sentiment analysis), which is an NLP task that combines two well-defined NLP sub-tasks, namely, Named Entity Recognition and Sentiment Analysis (Schouten et al., 2016; Nazir et al., 2020; Do et al., 2019). In this task, a machine-learning model is trained to detect entities (i.e., topics) within a piece of text and quantify the direction and strength of the sentiment expressed toward each entity. As defined by Google's Natural Language API, an entity "represents a phrase in the text that is a known entity, such as a person, an organization, or location".[14]

For each entity in the text, we observe three attributes: salience, sentiment score, and magnitude score. Salience, which ranges from 0 to 1, quantifies the importance of the entity in the text. More important entities receive a higher salience score. The sentiment score, which ranges from -1 to +1, describes the overall emotional leaning toward an entity. Scores between 0 and 1 indicate that the sentiment is positive, while scores between 0 and -1 indicate a negative sentiment. The sentiment score can also capture the strength of the emotion; the closer the score is to 1 (-1), the more positive (negative) the sentiment is. Scores closer to 0 can mean that either the sentiment towards the entity is not intense, or that both positive and negative sentiments are present in the text, thus canceling each other out. The magnitude score, which ranges from 0 to ∞, measures the overall strength of the sentiment towards the entity, regardless of its direction. Figure 6 provides examples

[14] For a more detailed definition of "entity", see `https://cloud.google.com/natural-language/docs/reference/rest/v1/Entity`, accessed January 2023

of the classification of entities and sentiment in our data, and Figure 7 displays the histograms of the sentiment and magnitude scores of the entities in our sample. Entities with higher magnitude scores also tend to have sentiment scores closer to -1 or 1; in our sample, the correlation between the magnitude score and the absolute value of the sentiment score is 0.77.[15] We refer to entities that have a higher magnitude score or a higher absolute value of the sentiment score as more intense. We use the absolute value of the sentiment score as the main measure of intensity and the magnitude score as a robustness check.

Figure 6: Examples of the Entity and Sentiment Classification



I do not like Eren

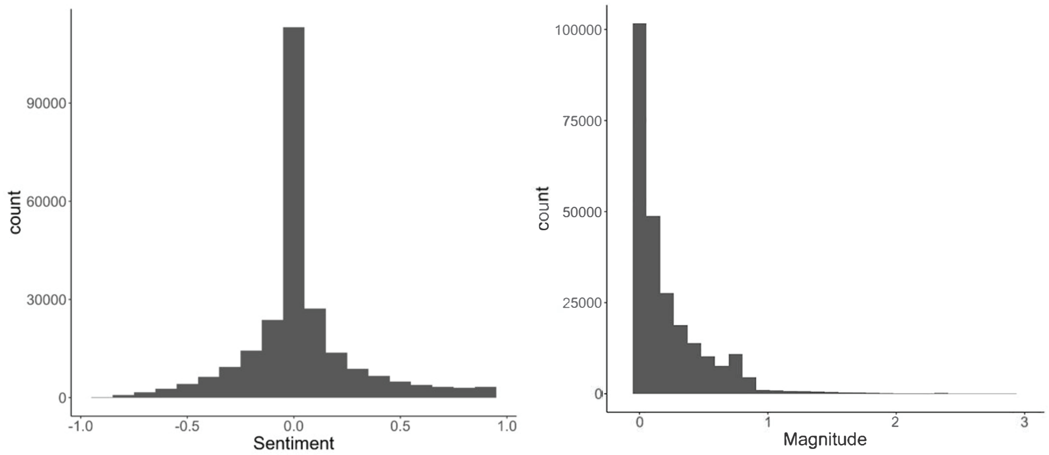
America is a failed state and it is only going to get worse

Most of the time, parenting is very rewarding

I like Kakashi. He is so positive despite everything he's been through.

🟦 Entity detected 🟩 Positive sentiment expressed towards entity 🟥 Negative sentiment expressed towards entity

Figure 7: Histograms of Sentiment and Magnitude Scores



(a) Distribution of Sentiment Scores

(b) Distribution of Magnitude Scores

These histograms plot distribution of the sentiment and magnitude scores for all entity mentions in our corpus with a salience score of greater than 0.1. Each observation is an entity.

[15] There may be instances in which the sentiment score is 0 but the magnitude score is not. In cases where the sentiment score is 0, the magnitude score can help distinguish between neutral content and conflicting emotions: a sentiment score of 0 and a magnitude close to 0 indicates that the text does not contain strong language, while a sentiment score of 0 and a high magnitude indicates the presence of both positive and negative opinions.

4.2.2 Correlation between Feedback and Intensity

Recall that we seek to understand how feedback towards a comment impacts the intensity in subsequent comments that mention the same entities as in the focal comment. Below, we describe how we construct our variables of interest before presenting the correlations between feedback towards the focal comment and the intensity of the "next mentions".

Let c denote the original comment, and $t(c)$ denotes the day that c is created. $E_c = \{e_1, e_2, ..., e_E\}$ is the set of entities in c, and let $\tilde{E}_c \subset E_c$ be a set of all "salient" entities in E_c. We define an entity to be "salient" if the entity's salience is greater than or equal to 0.1.[16] We include this restriction on salience because feedback is more likely to be directed at the more salient entities in a comment. We denote s_{ec} and m_{ec} to be the sentiment score and magnitude score of entity e in comment c, respectively.

The outcome of interest is the intensity of the "next mention" of salient entities in c. We illustrate how we define a "next mention" with the following example. Consider a comment c created by user i at day t with two salient entities: "A" and "B". Suppose user i creates another comment c' on day $t + 2$ that mentions either "A" or "B". Furthermore, let day $t + 2$ be the first day after t that user i comments about "A" or "B", so that c' is the first next mention of "A" or "B". If comment c' mentions only "A", then the outcome of interest is the intensity (i.e., absolute value of the sentiment score) towards entity "A" in comment c'. If c' mentions only "B", then the outcome of interest is the intensity towards entity "B" in comment c'. If c' mentions both "A" and "B", then the outcome of interest is the average of the intensity towards "A" and "B". Alternatively, if user i creates another comment c'' on the same day that mention "A" or "B", then we take the average of intensities towards "A" and "B" across all mentions in comments c' and c''.

More generally, the outcomes of interest are

$$\bar{s}'_{i,c} = \frac{1}{N} \sum_{c' \in C'} \sum_{e \in E_{c'}} \mathbb{1}\{e \in \tilde{E}_c\} \times s_{e,c'} \tag{3}$$

$$\bar{m}'_{i,c} = \frac{1}{N} \sum_{c' \in C'} \sum_{e \in E_{c'}} \mathbb{1}\{e \in \tilde{E}_c\} \times m_{e,c'}, , \tag{4}$$

where

$$N = \sum_{c' \in C'} \sum_{e \in E_{c'}} \mathbb{1}\{e \in \tilde{E}_c\}, \tag{5}$$

and C' is the set of comments that are (1) posted on day t' such that t' is the first day after $t(c)$

[16] The median salience across all entities in our sample is 0.05. The mean is 0.17. We test for robustness to alternative thresholds in Table A.1 and A.2 in the Appendix.

that i mentions any of the entities in the original comment c, and (2) contain at least one entity in $\tilde{E}_c$. In words, C' is the set of comments that contain the first "next mention(s)" of the salient entities in the original comment. Because 99.8% of all feedback is received within the first three days of posting, we restrict our main analysis to "next mentions" that occur within three days of the original comment's creation.[17]

To measure the correlation between feedback and subsequent content intensity, we estimate the following linear regression:

$$\begin{aligned}|\bar{s'}_{i,c}|= {} & \alpha_i + \alpha_{t(c')} + \alpha_{\tau(c')} + \alpha_{r(c)} \\ & + \beta_1 \mathbb{1}\{NegativeFeedback_{ic\tau}\} + \beta_2(\mathbb{1}\{NegativeFeedback_{ic\tau}\} \times |s_{ic}|) \\ & + \gamma_1 \mathbb{1}\{PositiveFeedback_{ic\tau}\} + \gamma_2(\mathbb{1}\{PositiveFeedback_{ic\tau}\} \times |s_{ic}|) \\ & + X'\Delta + \epsilon_{i,c,\tau}, \quad (6)\end{aligned}$$

In this specification, i indexes the user, c indexes the original comment. The outcome of interest, $|\bar{s'}_{i,c}|$, is the absolute value of the average sentiment score of the "next mention" of any of the entities in the original comment c by user i, as described in Equation 3. Recall that we take the absolute value because we are interested in the intensity of the sentiment, and not the direction. Let τ denotes the number of days elapsed between the original comment and the first next mention. For example, if comment c is created at t and the first day that user i creates another comment that mentions any of the entities in c is $t+2$, then $\tau = 2$. $\mathbb{1}\{NegativeFeedback_{ic\tau}\}$ is an indicator for whether comment c receives negative feedback τ days after c is created; this term equals 1 if $Score_{i,c,\tau} < Score_{i,c,\tau-1}$. Similarly, $\mathbb{1}\{PositiveFeedback_{ic\tau}\} = 1$ if $Score_{i,c,\tau} > Score_{i,c,\tau-1}$.

β_1 and γ_1 indicate the correlation between negative and positive feedback, respectively, and intensity. Furthermore, we expect that the extent of moderation depends on the intensity of the original comment. Therefore, we include the interaction between the indicators of positive and negative feedback with the average intensity of the original comment $|s_{ic}|$[18]; β_2 and γ_2 indicate the extent that the intensity of the original comment is associated with the effect of negative and positive feedback, respectively. Recall that a comment's score is measured with error due to Reddit's score "fuzzing". We assume that the error is random, which attentuates the estimated effects. Therefore,

[17] We test for robustness to alternative time windows in Table A.3 and A.4 in Appendix A.

[18] s_{ic} is the average of the sentiment scores of all salient entities in comment c, weighted by the entity's salience:

$$s_{ic} = \sum_{e \in \tilde{E}_c} (s_{e,c} \times \tilde{w}_{e,c}),$$

where $\tilde{w}_{e,c}$ is the reweighted salience of all entities in $\tilde{E}_c$ such that $\sum_{e \in \tilde{E}_c} \tilde{w}_{e,c} = 1$.

our coefficient estimates can be interpreted as lower bounds on the true effects.

In a similar spirit to Equation 2, we include a series of fixed effects to control for various unobserved characteristics that may induce a correlation between feedback and the intensity of subsequent posts. In particular, we include fixed effects for each user i (because users who post more intense content may get more feedback), and the calendar date of the "next mention" (because the intensity of content may vary across days). In addition, we include fixed effects for τ, the number of days elapsed between the day the original comment c is created and the first next mention (because intensity towards an entity may change over time in a nonlinear fashion), and $r(c)$, the subreddit of the original comment (because different subreddits may inherently vary in the intensity of content). Like Equation 2, we also control for the original comment's score on the previous day ($Score_{i,c,\tau-1}$, $Score^2_{i,c,\tau-1}$) and the absolute value of the amount of feedback the original comment received on day τ ($|Feedback_{i,c,\tau}|$, $Feedback^2_{i,c,\tau}$) in matrix $X_{ic\tau}$. The matrix X also includes $|s_{ic}|$, the intensity of the original comment, averaged over all salient entities. This variable controls for correlations in intensity across mentions (i.e., a user who is intense about topic A may also be more intense the next time he mentions A).

Table 7 reports the OLS estimates of Equation 6. We find that the main effects of negative and positive feedback (coefficients of $\mathbb{1}\{NegativeFeedback\}$ and $\mathbb{1}\{PositiveFeedback\}$) are not statistically significant. However, the interaction between negative feedback and the sentiment of the original comment (coefficient of $\mathbb{1}\{NegativeFeedback\} \times |s_c|$) is negative and statistically significant. This suggests that negative feedback is associated with moderation of intensity in subsequent comments if the original comment was more extreme. We see no such moderation for positive feedback.

4.2.3 Identification Strategy

Although Equation 6 controls for obvious sources of endogeneity (e.g., individual and time-invariant heterogeneity), there still may be other omitted variables that are correlated with feedback and the intensity of subsequent posts. A case in point is the previously mentioned Roe v. Wade example wherein the unobserved increased salience of the specific topic of abortion rights would impact both the direction of the feedback and the language used towards that topic in future mentions. The strategy that we employ to identify the causal relationship between feedback and subsequent intensity of posts is similar to the approach used earlier in Section 4.1.2; we rely on the idea that changes in the score of a post (the difference in upvotes and downvotes) are more salient once the score crosses a certain threshold. Recall that we conjecture that feedback is likely to impact future mentions of the entity (or entities) that the feedback is directed towards. Therefore, our unit of

Table 7: Correlation between Feedback and Intensity of Subsequent Content

Dependent Variables:	|Sentiment| of Next Mention
Model:	(1)
$\|s_c\|$	0.1208***
	(0.0127)
$\mathbb{1}\{NegativeFeedback\}$	0.0071
	(0.0046)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0700***
	(0.0232)
$\mathbb{1}\{PositiveFeedback\}$	0.0052
	(0.0036)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0283
	(0.0183)
User fixed effects	✓
Subreddit fixed effects	✓
Days After Posting fixed effects	✓
Date fixed effects	✓
Controls	✓
Observations	31,707
R^2	0.25962

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table presents the OLS estimates of Equation 6. Standard errors are clustered at the user level. Coefficient estimates of the score and feedback are omitted for readability. All coefficient estimates are presented in Table C.6 in the Appendix.

observation is an individual comment. However, unlike karma, the scores of individual comments tend to be significantly lower in magnitude; in our data, 88% of comments have a score between -10 and 10. This implies that we cannot use the left digit as the threshold at which feedback becomes more salient. As a result, we utilize a sign change as the threshold. In particular, holding the amount of feedback constant, changes in score are likely to be more salient when the sign changes. For example, we assume that negative feedback is more salient when a comment's score drops from 2 to -1 than when its score drops from 4 to 1, despite both changes in score representing a feedback of -3. We conduct a placebo test to check the validity of this assumption that feedback that is associated with a sign change will be more salient, which we discuss in Section 4.2.4.

To operationalize this identification strategy, we estimate the following linear regression:

$$\begin{aligned}
|\bar{s'}_{i,c}| = {} & \alpha_i + \alpha_{t(c')} + \alpha_\tau + \alpha_{r(c)} \\
& + \beta_1 \mathbb{1}\{NegativeFeedback_{ic\tau}\} + \beta_2(\mathbb{1}\{NegativeFeedback_{ic\tau}\} \times |s_{ic}|) \\
& + \beta_3 \mathbb{1}\{ScoreBecameNegative_{ic\tau}\} + \beta_4(\mathbb{1}\{ScoreBecameNegative_{ic\tau}\} \times |s_{ic}|) \\
& + \gamma_1 \mathbb{1}\{PositiveFeedback_{ic\tau}\} + \gamma_2(\mathbb{1}\{PositiveFeedback_{ic\tau}\} \times |s_{ic}|) \\
& + \gamma_3 \mathbb{1}\{ScoreBecamePositive_{ic\tau}\} + \gamma_4(\mathbb{1}\{ScoreBecamePositive_{ic\tau}\} \times |s_{ic}|) \\
& + X'\Delta + \epsilon_{ic\tau}, \quad (7)
\end{aligned}$$

This regression is identical to Equation 6, with the exception of four additional variables: $\mathbb{1}\{ScoreBecameNegative_{ic\tau}\}$, $\mathbb{1}\{ScoreBecamePositive_{ic\tau}\}$, and their interactions with $|s_{ic}|$. $\mathbb{1}\{ScoreBecameNegative_{ic\tau}\}$ and $\mathbb{1}\{ScoreBecamePositive_{ic\tau}\}$ are indicators for whether the score of the original comment changes from greater than or equal to 0 to less than 0 or changes from less than 0 to greater than or equal to 0 from day $\tau - 1$ to day τ, respectively. Given our identification strategy, the coefficients of interest are β_3 and β_4, which correspond to negative feedback, and γ_3 and γ_4, which correspond to positive feedback.

4.2.4 Results

We report the estimates of Equation 7 in column 1 of Table 8. Recall that we estimate the main effect of feedback ($\mathbb{1}\{ScoreBecameNegative\}$, $\mathbb{1}\{ScoreBecamePositive\}$) as well as their corresponding interactions with the magnitude of sentiment of the original comment ($|s_{ic}|$). Therefore, the main effects should be interpreted as the effect of feedback when the original comment had an intensity score of zero. Our results in Table 8 show that the main effect of $\mathbb{1}\{ScoreBecameNegative\}$ is positive and statistically significant. This means that when the original comment had an intensity score of zero, negative feedback increases the intensity of the next mention. However, the negative coefficient of the interaction between $\mathbb{1}\{ScoreBecameNegative\}$ and $|s_{ic}|$ means that negative feedback causes moderation when the original comment is intense (i.e., has sentiment close to -1 or 1). However, we do not find such an effect for positive feedback that is accompanied by a sign change (i.e., the coefficient of interaction between $\mathbb{1}\{ScoreBecamePositive\}$ and $|s_{ic}|$ is statistically indistinguishable from zero).

Table 8: Effect of Feedback on Intensity of Subsequent Content

	|Sentiment| of Next Mention
	(1)
$\mathbb{1}\{NegativeFeedback\}$	0.0031
	(0.0046)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0718**
	(0.0289)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0609***
	(0.0233)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.2645***
	(0.0939)
$\mathbb{1}\{PositiveFeedback\}$	0.0050
	(0.0036)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0049
	(0.0170)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0275
	(0.0183)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.0641
	(0.0979)
$\|s_c\|$	0.1207***
	(0.0127)
Observations	31,707
R^2	0.260
User fixed effects	✓
Subreddit fixed effects	✓
Days After Posting fixed effects	✓
Date fixed effects	✓
Controls	✓
✓	

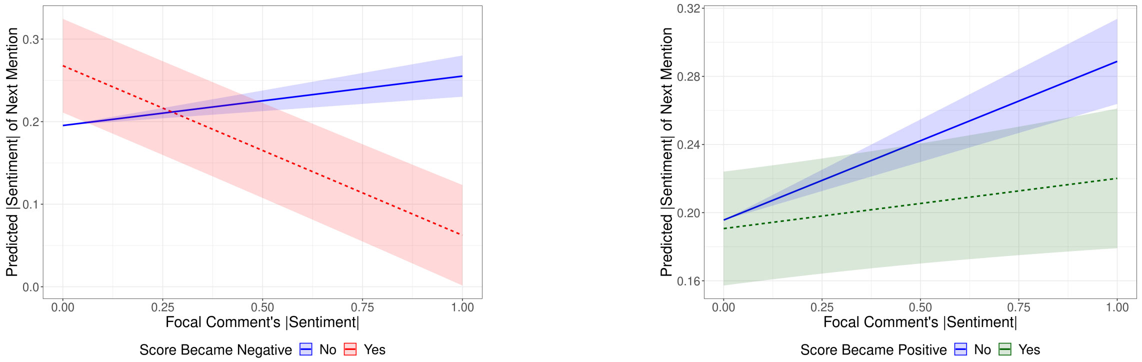
Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

OLS estimates of Equation 7. Standard errors are clustered at the user level. Coefficient estimates of the score and feedback are omitted for readability. All coefficient estimates are presented in column 1 of Table C.7 in the Appendix.

In Figure 8, we present a visual representation of the effect of feedback based on the estimates in Table 8. This figure plots the predicted intensity (i.e., the absolute value of the sentiment score,

$|s'_c|$) of the next mention as a function of the intensity of the original comment ($|s_c|$). In each panel, we hold the amount and direction of the feedback fixed; the solid line represents the predicted intensity of the next mention when the comment's score *does not* change signs, while the dashed line represents the predicted intensity when comment's score changes signs. Panels (a) and (b) of Figure 8 represent this for negative and positive feedback, respectively. Non-overlapping confidence intervals between the two lines means that there is a statistically significant difference in intensity of the subsequent content when feedback is more salient (i.e. the sign of the score changes).

Figure 8: Effect of Sign Change on Intensity: Predicted Absolute Sentiment Score



(a) Negative Feedback

(b) Positive Feedback

This figure plots the predicted absolute sentiment score of the next mention for when a user receives negative feedback (a) and positive feedback (b), assuming the user receives feedback of -1 (a) and 1 (b). The x-axis is the intensity of the original comment. The predicted values are obtained from the estimates from Table 8. Shaded regions represent the 95% confidence intervals.

In Figure 8a, we see that when the intensity of the original comment is zero, the confidence bands of the solid and the dashed lines do not overlap and the dashed line is above the solid line. This is consistent with the positive main effect of $\mathbb{1}\{ScoreBecameNegative\}$ and suggests that salient negative feedback (i.e., those associated with a sign change of the score) increases the intensity of the subsequent comment. However, as intensity of the original comment increases (to greater than 0.27), the countervailing negative effect of the interaction between $\mathbb{1}\{ScoreBecameNegative\}$ and $|s_c|$ moderates the intensity of subsequent posts. To put this in perspective, about 21% of the comments in our sample have an average absolute sentiment score greater than 0.27. Therefore, this moderating effect of negative feedback is reasonably common in our sample.

In Figure 8b, we present the corresponding predictions for positive feedback. The figure shows that the confidence intervals for the solid and dashed lines overlap except in the far right of the figure ($|s_c| > 0.88$). Therefore, there is no statistically significant effect of more salient positive feedback unless the original comment was extremely intense. In such cases, positive feedback moderates the intensity of subsequent content. However, less than 1% of comments in our sample exhibit this level of extreme intensity. Therefore, the moderating effect of positive feedback is rare in our sample.

We conduct a placebo check to verify whether our assumption—that a change in score from greater than or equal to zero to less than zero (or vice versa) is more salient than when the sign does not change, controlling for the amount of feedback — holds. We expect that the effect of feedback is larger when the score crosses the zero threshold as opposed to another threshold, say, 1. To test this, we redefined the variable $\mathbb{1}\{ScoreBecameNegative\}$ to equal 1 when the comment's score drops from greater than or equal to x to less than x.[19] We estimate Equation 7 using this new definition for $\mathbb{1}\{ScoreBecameNegative\}$ and $\mathbb{1}\{ScoreBecamePositive\}$ for other xs. The placebo test results are reported in Table 9. We find that the effects of interest become closer to zero and statistically insignificant when the thresholds are 1 and -1, as opposed to 0.

4.2.5 Robustness checks

We perform several robustness checks to verify that the results documented thus far are not an artifact of the way we specify our model or how we operationalize the various variables.

The first potential concern with the above analysis is that a sign change in scores is more likely to occur when the magnitude of feedback is large. Recall that we treat sign change in scores as an exclusion restriction to identify the effect of feedback on the intensity of subsequent posts. However, the magnitude of the feedback can directly influence the intensity of subsequent posts, thereby threatening our identification strategy. In our specification in Equation 7, we address for this concern by including a parametric function of the magnitude of the feedback as a control. However, it is still conceivable that the parametric specification does not fully capture the direct relationship between the magnitude of feedback and the intensity of subsequent posts. To account for this, we consider a subsample of comments that have a score of between -10 and 10 prior to the next mention. These comments are likely to be more comparable to each other once we include all the controls. The estimated coefficients for this subsample, which are reported in second column of Table C.7 in the Appendix, are similar to those presented of the full sample (Table 8), suggesting that our estimates are not driven by unobserved factors that correlate with large amounts of feedback.

A second concern is that our measured effects are because of the absolute value of the sentiment score being bounded between 0 and 1. As a result, moderation is more likely for extreme sentiments (absolute sentiment scores closer to 1) than for moderate sentiments (sentiment scores closer to 0)

[19] Similarly, $\mathbb{1}\{ScoreBecamePositive\} = 1$ when the score increases from less than or equal to x to greater than x.

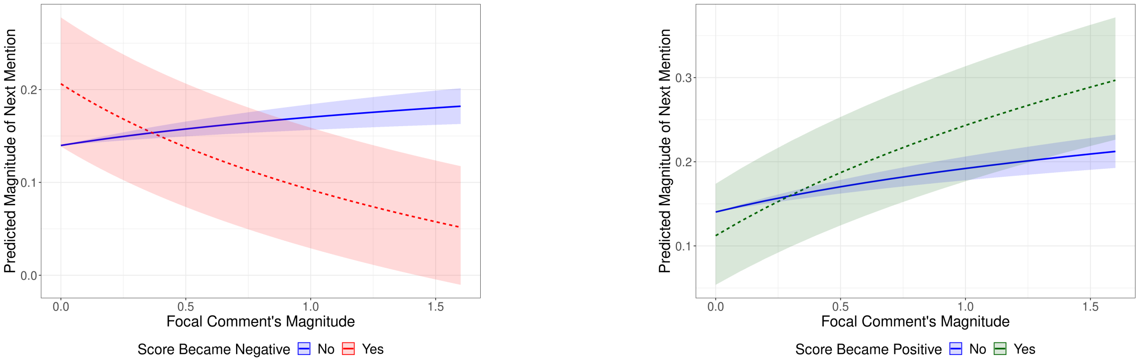
Table 9: Effects of Feedback on Intensity: Placebo Tests

	|Sentiment| of Next Mention	
	(1)	(2)
$\|s_c\|$	0.1208***	0.1208***
	(0.0127)	(0.0127)
$\mathbb{1}\{NegativeFeedback\}$	0.0065	0.0065
	(0.0045)	(0.0047)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0245	0.0159
	(0.0353)	(0.0237)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0666***	-0.0693***
	(0.0228)	(0.0238)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.1286	-0.0199
	(0.1777)	(0.1194)
$\mathbb{1}\{PositiveFeedback\}$	0.0054	0.0055
	(0.0036)	(0.0038)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0140	-0.0034
	(0.0187)	(0.0121)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0289	-0.0313*
	(0.0184)	(0.0186)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	0.0563	0.0345
	(0.1027)	(0.0651)
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Controls	✓	✓
Sign Change Threshold	-1	1
Observations	31,707	31,707
R^2	0.25967	0.25966

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

OLS estimates of Equation 7. $ScoreBecameNegative = 1$ if the score drops from greater than or equal to x to less than x, where x is the sign change threshold.$ScoreBecamePositive = 1$ if the score increases from less than x to greater than or equal to x. Column 1 reports the estimates when $x = 1$, and column 2, $x = -1$. Estimated coefficients for score and feedback are omitted for readability. Standard errors are clustered at the user level.

Figure 9: Effect of Sign Change on Intensity: Predicted Log Magnitude



(a) Negative Feedback

(b) Positive Feedback

This figure plots the predicted logged magnitude score of the next mention for when a user receives negative feedback (a) and positive feedback (b), assuming the user receives feedback of -1 (a) and 1 (b). The predicted values are obtained from the estimates from Table C.8 in the Appendix. Shaded regions represent the 95% confidence intervals.

simply due to the upper bound. To address this concern, we measure the effects of feedback on the (logged) magnitude score, which is unbounded from above. We estimate Equation 7 but instead of the average absolute sentiment score, the dependent variable is the logged average magnitude score of the next mentions of the entities in the focal comment, $log(\bar{m}'_{ic})$, where $\bar{m}'_{ic}$ is defined in Equation 4. In addition, we interact the feedback and score change indicators with the magnitude of the focal comment $\log(m_{ic})$, rather than $|s_{ic}|$. Figure 9 plots the predicted logged magnitude score of the next mentions of the entities in the focal comment. We report the corresponding coefficient estimates in Table C.8 in the Appendix. We find similar results: more salient negative feedback moderates the intensity of extreme content relative to less salient negative feedback (Figure 9a). We find no such intensity moderation effects of more salient positive feedback (Figure 9b).

We also conduct the following robustness checks around sample definitions. Recall that our intensity analysis focused on “salient” entities—entities with salience scores of greater than 0.1. The idea behind this approach is that a comment might refer to multiple entities, some of which are more salient than others. Since our dependent variable is the intensity of the subsequent comment when the same entity is mentioned, considering entities that were not important in the original comment is likely to introduce noise. Although restricting our analysis to entities that have a minimum salience score can reduce this noise, our results can also be sensitive to the choice of this threshold. To verify whether this is the case, we estimate our model using alternative thresholds for salience: 0.05 (the median), and 0.15. We find that the qualitative results are robust. The results are displayed in

Tables A.1 and A.2 in the Appendix.

We had also limited the sample to “next mentions” within three days of focal comment creation because 99.8% of all feedback is received within this time window. We verify the robustness of our results to this restriction by considering alternative time windows of two, four, and five days after the focal comment was created. The results in Tables A.3 and A.4 in the Appendix show that the substantive insights remain unaltered when we consider alternative time windows. The effects are also robust to alternative controls, such as the logged magnitude of feedback and score, and the final score of the comment (as a proxy for comment quality). Section A in the Appendix details our robustness checks.

5 Discussion

Peer feedback systems are a prevalent feature on social media platforms, yet our understanding of their influence on content production remains incomplete. This knowledge gap presents an exciting opportunity for future research. In this study, we investigate how receiving negative peer feedback affects a user’s subsequent content production. We find that receiving negative peer feedback causes users to produce more content. In addition, negative feedback encourages users to participate in communities they had not participated in recently. On the other hand, we do not find evidence that negative feedback deters individuals from participating in communities in which they receive the negative feedback. We also find that receiving negative peer feedback changes the nature of subsequent content; negative peer feedback moderates the intensity of subsequent content, especially when the original comment that received the negative feedback is intense.

There are two potential, somewhat related, explanations for these effects. The first is reputation management, wherein users respond to feedback in ways that, they believe, will enhance their standing on the platform. This explanation can rationalize why users tend to produce more content upon receiving both positive and negative feedback. Furthermore, reputation management can explain why users moderate their extreme views when they receive negative feedback. The second explanation is that users are learning about what types of content to post in order to generate positive feedback and then respond appropriately to negative feedback. In some sense, the objective function that the user is attempting to optimize might still be their standing on the platform.

Despite having similar objectives, these two mechanisms present nuanced implications regarding the responsiveness of different user types to negative feedback. The learning mechanism suggests that the impact of feedback would be greater for less experienced users (i.e., users with fewer posts). In contrast, the reputation management mechanism suggests that both inexperienced and experienced users may respond to feedback, albeit for different reasons.

Inexperienced users may have worse reputations (lower karma) simply because they do not have much posting experience, which implies that a reduction in karma has a larger proportional impact on their overall reputation. Consequently, a decrease in karma would have a larger proportional impact on their overall reputation. Therefore, we might anticipate that the effects of negative feedback are pronounced among less experienced users. Conversely, more experienced users, who have a greater posting history, may prioritize maintaining their reputation more than less experienced users. Consequently, they are likely to be more motivated to rebuild their reputation following a decline in karma. Thus, the effects of negative feedback may also be pronounced among experienced users due to their heightened concern for reputation management.

Motivated by these two potential mechanisms and their implications for inexperienced versus experienced users, we explore the heterogeneity of the negative feedback effects by user experience, measured by the number of comments they created prior to data collection.[20] We do not find evidence that the effects of (salient) negative feedback on both incidence and intensity are different by user experience.[21] Therefore, our data are more consistent with the reputation management mechanism than learning, which predicts that effects of feedback are larger for inexperienced users.

Our paper has implications for social media platforms on several fronts. First, our findings that negative feedback can increase content production and lead to moderation of tone can be an encouraging sign regarding the potential benefits of allowing negative feedback. Second, we had conjectured that negative feedback might deter users from creating content. Consequently, the space would be populated only with users expressing the accepted opinions, thereby facilitating the formation of echo chambers. At the same time, it was unclear whether negative feedback would foster moderation of tone or drive users to harden their positions with more extreme positions. Of these, the latter would have resulted in a situation where negative feedback increased polarization of content. Our finding that negative feedback encourages users to create more content and moderate their extreme viewpoints suggests that “downvotes”, “dislikes”, etc., can serve as the whip that regulates the polarization of content on online platforms. On the other hand, this raises a philosophical question of whether a reduction in polarization (i.e., all individuals moving towards a moderate opinion) is another form of an echo chamber. Elaborating on this debate is beyond the scope of this paper but would be an interesting avenue for future research.

[20] The number of prior comments range from 0 to 124,971, with a median of 216 comments.

[21] Tables B.1 and B.2 in the Appendix reports the effects on incidence and intensity, respectively.

6 Conclusion

Our study suffers from a few limitations that also highlight potential avenues for future research. First, we have only considered quantitative feedback provided to the users in the form of upvotes/downvotes. Future research can explore the impact of qualitative feedback provided to the users in the form of comments and complex interactions between quantitative and qualitative forms of feedback. Second, we do not have information on how much individuals care about the entities for which they express an opinion. A measure that quantifies how much individuals care about entities can help in further improving our understanding of how feedback systems alter the nature of subsequent content. Third, although our effects are consistent with reputation management as the underlying mechanism, we are unable to pin down whether this is definitively the case. To do so, we need data from a setting in which there is variation in whether reputation is formed or observed. An example would be Reddit implementing a change to reduce the visibility of karma to other users. Relatedly, our findings are measured in a context in which an objective measure of reputation is visible. Investigating how negative feedback impacts UGC in other contexts in which such information is received in different ways holds potential for future research. Finally, it is an empirical question as to whether our findings about negative feedback fostering more content production and moderating the tone would generalize to emotionally charged contexts such as politics, religion, and even sports. Future research can potentially investigate this using data from some of these contexts. Despite these limitations, we hope that our research advances our understanding of the role of peer feedback systems in driving content production on social media platforms and stimulates additional research in this area.

References

Anderson, Eric T and Simester, Duncan I (2003). "Effects of $9 price endings on retail sales: Evidence from field experiments". Quantitative marketing and Economics, 1 (1), 93–110.

Berger, Jonah, Sorensen, Alan T, and Rasmussen, Scott J (2010). "Positive effects of negative publicity: When negative reviews increase sales". Marketing science, 29 (5), 815–827.

Bizer, George Y and Schindler, Robert M (2005). "Direct evidence of ending-digit drop-off in price information processing". Psychology & Marketing, 22 (10), 771–783.

Burtch, Gordon, He, Qinglai, Hong, Yili, and Lee, Dokyun (2022). "How do peer awards motivate creative content? Experimental evidence from Reddit". Management Science, 68 (5), 3488–3506.

Cheng, Justin, Danescu-Niculescu-Mizil, Cristian, and Leskovec, Jure (2014). "How community feedback shapes user behavior". In "Proceedings of the International AAAI Conference on Web and Social Media", volume 8, 41–50.

Cinelli, Matteo, De Francisci Morales, Gianmarco, Galeazzi, Alessandro, Quattrociocchi, Walter, and Starnini, Michele (2021). "The echo chamber effect on social media". Proceedings of the National Academy of Sciences, 118 (9), e2023301118.

Dandekar, Pranav, Goel, Ashish, and Lee, David T (2013). "Biased assimilation, homophily, and the dynamics of polarization". Proceedings of the National Academy of Sciences, 110 (15), 5791–5796.

Deci, Edward L and Cascio, Wayne F (1972). "Changes in intrinsic motivation as a function of negative feedback and threats."

Denny, Paul (2013). "The effect of virtual achievements on student engagement". In "Proceedings of the SIGCHI conference on human factors in computing systems", 763–772.

Do, Hai Ha, Prasad, Penatiyana WC, Maag, Angelika, and Alsadoon, Abeer (2019). "Deep learning for aspect-based sentiment analysis: a comparative review". Expert systems with applications, 118, 272–299.

Eckles, Dean, Kizilcec, René F, and Bakshy, Eytan (2016). "Estimating peer effects in networks with peer encouragement designs". Proceedings of the National Academy of Sciences, 113 (27), 7316–7322.

Flaxman, Seth, Goel, Sharad, and Rao, Justin M (2016). "Filter bubbles, echo chambers, and online news consumption". Public opinion quarterly, 80 (S1), 298–320.

Fong, Jessica and Hunter, Megan (2022). "Can facing the truth improve outcomes? effects of information in consumer finance". Marketing Science, 41 (1), 33–50.

Freedman, Jonathan L and Sears, David O (1965). "Selective exposure". In "Advances in experimental social psychology", volume 2, 57–97. Elsevier.

Frey, Dieter (1986). "Recent research on selective exposure to information". Advances in experimental social psychology, 19, 41–80.

Gallus, Jana (2017). "Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia". Management Science, 63 (12), 3999–4015.

Huang, Justin T, Kaul, Rupali, and Narayanan, Sridhar (2022). "The Causal Effect of Attention and Recognition on the Nature of User-Generated Content: Experimental Results from an Image-Sharing Social Network".

Huang, Justin T and Narayanan, Sridhar (2020). "Effects of Attention and Recognition on Engagement, Content Creation and Sharing: Experimental Evidence from an Image Sharing Social Network".

Husain, S Ali, King, Kristen L, and Mohan, Sumit (2021). "Left-digit bias and deceased donor kidney utilization". Clinical transplantation, 35 (6), e14284.

Jin, Jiahua, Li, Yijun, Zhong, Xiaojia, and Zhai, Li (2015). "Why users contribute knowledge to online communities: An empirical study of an online social Q&A community". Information & management, 52 (7), 840–849.

Kandel, Denise B (1978). "Similarity in real-life adolescent friendship pairs." Journal of personality and social psychology, 36 (3), 306.

Lacetera, Nicola, Pope, Devin G, and Sydnor, Justin R (2012). "Heuristic thinking and limited attention in the car market". American Economic Review, 102 (5), 2206–2236.

Levy, Ro'ee (2021). "Social media, news consumption, and polarization: Evidence from a field experiment". American economic review, 111 (3), 831–870.

List, John A, Muir, Ian, Pope, Devin, and Sun, Gregory (2023). "Left-Digit Bias at Lyft". Review of Economic Studies, rdad014.

Lu, Shijie, Xie, Ying, and Chen, Xingyu (2022). “Immediate and enduring effects of digital badges on online content consumption and generation”. International Journal of Research in Marketing.

McPherson, Miller, Smith-Lovin, Lynn, and Cook, James M (2001). “Birds of a feather: Homophily in social networks”. Annual review of sociology, 415–444.

Moe, Wendy W and Schweidel, David A (2012). “Online product opinions: Incidence, evaluation, and evolution”. Marketing Science, 31 (3), 372–386.

Mummalaneni, Simha, Yoganarasimhan, Hema, and Pathak, Varad V (2022). “Producer and Consumer Engagement on Social Media Platforms”. Working paper.

Nazir, Ambreen, Rao, Yuan, Wu, Lianwei, and Sun, Ling (2020). “Issues and challenges of aspect-based sentiment analysis: a comprehensive survey”. IEEE Transactions on Affective Computing.

Nussbaum, A David and Dweck, Carol S (2008). “Defensiveness versus remediation: Self-theories and modes of self-esteem maintenance”. Personality and Social Psychology Bulletin, 34 (5), 599–612.

Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry (1999). “The PageRank citation ranking: Bringing order to the web.” Technical report, Stanford InfoLab.

Pariser, Eli (2011). The filter bubble: What the Internet is hiding from you. penguin UK.

Peng, Jing, Agarwal, Ashish, Hosanagar, Kartik, and Iyengar, Raghuram (2018). “Network overlap and content sharing on social media platforms”. Journal of marketing research, 55 (4), 571–585.

Rafieian, Omid and Yoganarasimhan, Hema (2023). “AI and personalization”. Artificial Intelligence in Marketing, 77–102.

Schouten, Kim, Frasincar, Flavius, and Dekker, Rommert (2016). “An information gain-driven feature study for aspect-based sentiment analysis”. In “International Conference on Applications of Natural Language to Information Systems”, 48–59. Springer.

Shriver, Scott K, Nair, Harikesh S, and Hofstetter, Reto (2013). “Social ties and user-generated content: Evidence from an online social network”. Management Science, 59 (6), 1425–1443.

Stroud, Natalie Jomini (2008). “Media use and political predispositions: Revisiting the concept of selective exposure”. Political Behavior, 30 (3), 341–366.

Strulov-Shlain, Avner (2022). “More than a Penny’s Worth: Left-Digit Bias and Firm Pricing”. The Review of Economic Studies.

Sunstein, Cass R (2009). Going to extremes: How like minds unite and divide. Oxford University Press.

Thomas, Manoj and Morwitz, Vicki (2005). “Penny wise and pound foolish: the left-digit effect in price cognition”. Journal of Consumer Research, 32 (1), 54–64.

Toubia, Olivier and Stephen, Andrew T (2013). “Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter?” Marketing Science, 32 (3), 368–392.

Wang, Yao and Majeed, Abdul (2022). “How do users’ feedback influence creators’ contributions: an empirical study of an online music community”. Behaviour & Information Technology, 1–17.

Yoganarasimhan, Hema and Yakovetskaya, Irina (2022). “Is Social Media Seeded with Polarizing News?” Working paper.

Zhang, Xiaoquan and Zhu, Feng (2006). “Intrinsic motivation of open content contributors: The case of Wikipedia”. In “Workshop on information systems and economics”, volume 10. Citeseer.

Zhu, Kai, Khern-am nuai, Warut, and Yu, Yinan (2022). “Any Feedback is Welcome: Peer Feedback and User Behavior on Digital Platforms”. Working paper.

APPENDIX

A Additional Robustness Checks

We conduct robustness checks using both measures of intensity (magnitude m_c and sentiment s_c). The effects of feedback on intensity are robust to

- Alternative salience thresholds of 0.05 and 0.15 (Tables A.1 and A.2)
- Selecting only next mentions within 2, 4, or 5 days the focal comment is created (Tables A.3 and A.4)
- A log specification of total feedback and score (Column 1 of Tables A.5 and A.6)
- Controlling for the "final score" of the comment, in which the final score is the last observed score (Column 2 of Tables A.5 and A.6)

Table A.1: Robustness to Salience Thresholds: Sentiment

	|Sentiment| of Next Mention	
	(1)	(2)
$\mathbb{1}\{NegativeFeedback\}$	0.0060	0.0036
	(0.0043)	(0.0047)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0496^{**}	0.0725^{**}
	(0.0239)	(0.0300)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0637^{***}	-0.0465^{**}
	(0.0225)	(0.0235)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.1806^{**}	-0.2389^{***}
	(0.0735)	(0.0876)
$\mathbb{1}\{PositiveFeedback\}$	0.0036	0.0051
	(0.0034)	(0.0039)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0087	-0.0043
	(0.0169)	(0.0192)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0306^{*}	-0.0331^{*}
	(0.0172)	(0.0195)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.1016	0.0053
	(0.0871)	(0.1024)
$\|s_c\|$	0.1166^{***}	0.1241^{***}
	(0.0122)	(0.0134)
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Controls	✓	✓
Salience Threshold	0.05	0.15
Observations	36,499	27,819
R^2	0.24677	0.27213

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level.

Table A.2: Robustness to Salience Thresholds: Magnitude

	Magnitude of Next Mention (log)	
	(1)	(2)
$\mathbb{1}\{NegativeFeedback\}$	0.0137*	0.0102**
	(0.0072)	(0.0047)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0403*	0.0540*
	(0.0242)	(0.0305)
$\mathbb{1}\{NegativeFeedback\} \times \log(m_c)$	-0.0614*	-0.0524***
	(0.0352)	(0.0172)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(m_c)$	-0.1365*	-0.1710
	(0.0725)	(0.1040)
$\mathbb{1}\{PositiveFeedback\}$	0.0127*	0.0100**
	(0.0069)	(0.0042)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0168	-0.0368
	(0.0263)	(0.0302)
$\mathbb{1}\{PositiveFeedback\} \times \log(m_c)$	-0.0406	-0.0244
	(0.0331)	(0.0170)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(m_c)$	0.0240	0.2168
	(0.1351)	(0.2031)
$\log(m_c)$	0.1017***	0.0872***
	(0.0326)	(0.0094)
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Controls	✓	✓
Salience Threshold	0.05	0.15
Observations	36,499	27,819
R^2	0.21652	0.26506

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level.

Table A.3: Robustness to Time Window after Focal Comment is Created: Sentiment

	|Sentiment| of Next Mention		
	(1)	(2)	(3)
$\mathbb{1}\{NegativeFeedback\}$	0.0052	0.0039	0.0051
	(0.0050)	(0.0045)	(0.0044)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0560*	0.0731***	0.0622**
	(0.0308)	(0.0283)	(0.0274)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0582**	-0.0480**	-0.0457**
	(0.0248)	(0.0225)	(0.0217)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.2369**	-0.2638***	-0.2584***
	(0.1104)	(0.0857)	(0.0847)
$\mathbb{1}\{PositiveFeedback\}$	0.0025	0.0061*	0.0063*
	(0.0040)	(0.0034)	(0.0033)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0060	-0.0097	0.0031
	(0.0199)	(0.0169)	(0.0221)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0318	-0.0276	-0.0338*
	(0.0203)	(0.0183)	(0.0181)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.0468	-0.0298	-0.1047
	(0.1050)	(0.0978)	(0.1086)
$\|s_c\|$	0.1096***	0.1220***	0.1267***
	(0.0134)	(0.0123)	(0.0122)
User fixed effects	✓	✓	✓
Subreddit fixed effects	✓	✓	✓
Days After Posting fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Controls	✓	✓	✓
Days After Posting	2	4	5
Observations	26,571	35,272	38,108
R^2	0.27578	0.24618	0.23633

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level.

Table A.4: Robustness to Time Window after Focal Comment is Created: Magnitude

	Magnitude of Next Mention (log)		
	(1)	(2)	(3)
$\mathbb{1}\{NegativeFeedback\}$	0.0138***	0.0065	0.0064
	(0.0050)	(0.0045)	(0.0044)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0376	0.0596**	0.0490*
	(0.0308)	(0.0295)	(0.0278)
$\mathbb{1}\{NegativeFeedback\} \times \log(m_c)$	-0.0622***	-0.0326*	-0.0267
	(0.0174)	(0.0169)	(0.0166)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(m_c)$	-0.1110	-0.1888*	-0.1817**
	(0.0984)	(0.0982)	(0.0897)
$\mathbb{1}\{PositiveFeedback\}$	0.0075	0.0099**	0.0092**
	(0.0047)	(0.0039)	(0.0037)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0323	-0.0335	-0.0246
	(0.0308)	(0.0260)	(0.0287)
$\mathbb{1}\{PositiveFeedback\} \times \log(m_c)$	-0.0211	-0.0189	-0.0209
	(0.0188)	(0.0160)	(0.0149)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(m_c)$	0.1482	0.1298	0.0969
	(0.1824)	(0.1554)	(0.1563)
$\log(m_c)$	0.0804***	0.0842***	0.0859***
	(0.0102)	(0.0087)	(0.0086)
User fixed effects	✓	✓	✓
Subreddit fixed effects	✓	✓	✓
Days After Posting fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Controls	✓	✓	✓
Days After Posting	2	4	5
Observations	26,571	35,272	38,108
R^2	0.26557	0.23841	0.23065

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level.

Table A.5: Alternative Controls: Sentiment

	|Sentiment| of Next Mention	
	(1)	(2)
$\mathbb{1}\{NegativeFeedback\}$	0.0208*	0.0212*
	(0.0111)	(0.0111)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0726**	0.0737**
	(0.0292)	(0.0299)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0608***	-0.0602***
	(0.0233)	(0.0233)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.2680***	-0.2705***
	(0.0931)	(0.0944)
$\mathbb{1}\{PositiveFeedback\}$	0.0107	0.0100
	(0.0065)	(0.0065)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0085	-0.0039
	(0.0171)	(0.0172)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0273	-0.0269
	(0.0183)	(0.0183)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.0627	-0.0580
	(0.0969)	(0.0929)
$\|s_c\|$	0.1209***	0.1209***
	(0.0127)	(0.0127)
$\log(\|FinalScore\|)$		-0.0126**
		(0.0060)
$\mathbb{1}\{FinalScore < 0\}$		-0.0318
		(0.0205)
$\log(\|FinalScore\|) \times \mathbb{1}\{FinalScore < 0\}$		0.0477***
		(0.0162)
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
logarithmic Score, Feedback Controls	✓	✓
Observations	31,707	31,707
R^2	0.26005	0.26022

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Standard errors are clustered at the user level.

Table A.6: Alternative Controls: Magnitude

	Magnitude of Next Mention (log)	
	(1)	(2)
$\mathbb{1}\{NegativeFeedback\}$	0.0187*	0.0193*
	(0.0106)	(0.0106)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0578*	0.0578*
	(0.0295)	(0.0317)
$\mathbb{1}\{NegativeFeedback\} \times \log(m_c)$	-0.0460**	-0.0457**
	(0.0185)	(0.0185)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(m_c)$	-0.1798*	-0.1858*
	(0.0979)	(0.0987)
$\mathbb{1}\{PositiveFeedback\}$	0.0092	0.0090
	(0.0066)	(0.0066)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0263	-0.0238
	(0.0280)	(0.0281)
$\mathbb{1}\{PositiveFeedback\} \times \log(m_c)$	-0.0199	-0.0198
	(0.0162)	(0.0162)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(m_c)$	0.0972	0.0995
	(0.1672)	(0.1666)
$\log(m_c)$	0.0840***	0.0840***
	(0.0087)	(0.0087)
$\log(\|FinalScore\|)$		-0.0068
		(0.0079)
$\mathbb{1}\{FinalScore < 0\}$		-0.0258
		(0.0214)
$\log(\|FinalScore\|) \times \mathbb{1}\{FinalScore < 0\}$		0.0398**
		(0.0164)
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
logarithmic Score, Feedback Controls	✓	✓
Observations	31,707	31,707
R^2	0.25120	0.25130

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Standard errors are clustered at the user level.

B Heterogeneity by User Experience

Table B.1: Incidence: Heterogeneity by User Experience

	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	3.025*** (0.4143)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \log(NComments)$	-0.0633 (0.0641)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	2.358** (1.183)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\} \times \log(NComments)$	-0.3627 (0.2500)
$\mathbb{1}\{PositiveFeedback_{it}\}$	3.241*** (0.1731)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \log(NComments)$	-0.0700** (0.0277)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	1.013** (0.4254)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\} \times \log(NComments)$	-0.1062 (0.0892)
User fixed effects	✓
Date fixed effects	✓
Controls	✓
Observations	49,197
Pseudo R^2	0.49032

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

$NComments$ is the number of comments the user made prior to the data collection period. Standard errors are clustered at the user level. Coefficient estimates of the comment karma, feedback, and distance to the nearest threshold are omitted for readability.

Table B.2: Intensity: Heterogeneity by User Experience

	|Sentiment| of Next Mention (1)
$\mathbb{1}\{NegativeFeedback\}$	0.0226 (0.0191)
$\mathbb{1}\{ScoreBecameNegative\}$	0.2327** (0.1035)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0795 (0.1114)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.5185 (0.3908)
$\mathbb{1}\{NegativeFeedback\} \times \log(NComments)$	-0.0023 (0.0022)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(NComments)$	-0.0209 (0.0127)
$\mathbb{1}\{NegativeFeedback\} \times \log(NComments) \times \|s_c\|$	0.0029 (0.0128)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(NComments) \times \|s_c\|$	0.0338 (0.0441)
$\mathbb{1}\{PositiveFeedback\}$	0.0285* (0.0153)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0319 (0.0469)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.1468* (0.0789)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.2282 (0.4633)
$\mathbb{1}\{PositiveFeedback\} \times \log(NComments)$	-0.0030* (0.0018)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(NComments)$	0.0033 (0.0055)
$\mathbb{1}\{PositiveFeedback\} \times \times \log(NComments) \times \|s_c\|$	0.0156* (0.0093)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(NComments) \times \|s_c\|$	0.0222 (0.0612)
$\log(NComments)$	0.2431 (1,600.2)
$\log(NComments) \times \|s_c\|$	-0.0144** (0.0071)
User fixed effects	✓
Subreddit fixed effects	✓
Days After Posting fixed effects	✓
Date fixed effects	✓
Controls	✓
Observations	31,706
R^2	0.261

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

NComments is the number of comments the user made prior to the data collection period. Standard errors are clustered at the user level. Coefficient estimates of the score and feedback are omitted for readability.

C Additional Tables

Table C.1: Correlation between Feedback and Incidence: All coefficients

Dependent Variable:	$\mathbb{1}\{Comment_{it}\}$
Model:	(1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.697^{***}
	(0.0963)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.837^{***}
	(0.0502)
$\lvert Feedback_{it} \rvert$	0.0080^{***}
	(0.0027)
$Feedback_{it}^2$	$-1.23 \times 10^{-6}\ ^{**}$
	(5.11×10^{-7})
$\mathbb{1}\{Comment_{i,t-1}\}$	0.2191^{***}
	(0.0372)
$Karma_{i,t-1}$	-5.84×10^{-5}
	(7.27×10^{-5})
$Karma_{i,t-1}^2$	6.12×10^{-11}
	(3.8×10^{-11})
$DistToThreshold_{i,t-1}$	2.59×10^{-5}
	(7.67×10^{-5})
$DistToThreshold_{i,t-1}^2$	-7.57×10^{-11}
	(1.48×10^{-10})
User fixed effects	✓
Date fixed effects	✓
Observations	49,197
Pseudo R^2	0.48997

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All coefficients for the specifications reported in Table 2. Standard errors are clustered at the user level.

Table C.2: Effect of Feedback on Incidence: All Coefficients

Dependent Variable:		$\mathbb{1}\{Comment_{it}\}$	
Model:	(1)	(2)	(3)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.639^{***}	2.331^{***}	2.613^{***}
	(0.0990)	(0.1004)	(0.1529)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	1.169^{**}	0.9411^{*}	1.145^{**}
	(0.4941)	(0.5326)	(0.4992)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.805^{***}	2.228^{***}	2.811^{***}
	(0.0491)	(0.0503)	(0.0789)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.6512^{***}	0.4001^{**}	0.6923^{***}
	(0.1386)	(0.1624)	(0.2013)
$\|Feedback_{it}\|$	0.0073^{***}	0.1388^{***}	0.0105
	(0.0025)	(0.0091)	(0.0088)
$Feedback^2_{it}$	$-1.12 \times 10^{-6**}$	-0.0021^{***}	-3.77×10^{-6}
	(4.91×10^{-7})	(0.0002)	(3×10^{-6})
$\mathbb{1}\{Comment_{i,t-1}\}$	0.2247^{***}	0.2155^{***}	0.2338^{***}
	(0.0371)	(0.0374)	(0.0612)
$Karma_{i,t-1}$	-5.71×10^{-5}	-0.0001^{*}	-0.0007
	(7.2×10^{-5})	(8.62×10^{-5})	(0.0006)
$Karma^2_{i,t-1}$	5.71×10^{-11}	$6.6 \times 10^{-10*}$	1.44×10^{-7}
	(3.72×10^{-11})	(3.56×10^{-10})	(9.26×10^{-8})
$DistFromThreshold_{i,t-1}$	3.28×10^{-5}	2.21×10^{-5}	0.0066
	(7.58×10^{-5})	(0.0001)	(0.0097)
$DistFromThreshold^2_{i,t-1}$	-8.91×10^{-11}	$2.45 \times 10^{-9*}$	-0.0002
	(1.45×10^{-10})	(1.47×10^{-9})	(0.0002)
User fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Sample	All	$\|$Feedback$\| \leq 50$	DistFromThreshold ≤ 50
Observations	49,197	46,723	24,310
Pseudo R^2	0.49056	0.47023	0.41587

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All coefficients for the specifications reported in Table 3. Standard errors are clustered at the user level.

Table C.3: Placebo Test for Effect of Feedback on Incidence: All Coefficients

Dependent Variable: Model:	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.693^{***} (0.0976)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChangedPlacebo_{it}\}$	0.2648 (0.8369)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.834^{***} (0.0503)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChangedPlacebo_{it}\}$	0.3061 (0.3198)
$\lvert Feedback_{it}\rvert$	0.0080^{***} (0.0027)
$Feedback^2_{it}$	$-1.23 \times 10^{-6**}$ (5.12×10^{-7})
$\mathbb{1}\{Comment_{i,t-1}\}$	0.2181^{***} (0.0371)
$Karma_{i,t-1}$	-5.84×10^{-5} (7.27×10^{-5})
$Karma^2_{i,t-1}$	6.12×10^{-11} (3.8×10^{-11})
$DistFromThreshold_{i,t-1}$	2.65×10^{-5} (7.67×10^{-5})
$DistFromThreshold^2_{i,t-1}$	-7.6×10^{-11} (1.48×10^{-10})
User fixed effects	✓
Date fixed effects	✓
Observations	49,197
Pseudo R^2	0.48999

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All coefficients for the specification reported in Table 4. Standard errors are clustered at the user level.

Table C.4: Robustness with Fixed Effects for Number of Digits in Karma: All Coefficients

	$\mathbb{1}\{Comment_{it}\}$ (1)
$\mathbb{1}\{NegativeFeedback_{it}\}$	2.639^{***} (0.0990)
s $\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	1.190^{**} (0.4983)
$\mathbb{1}\{PositiveFeedback_{it}\}$	2.806^{***} (0.0491)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.6597^{***} (0.1405)
$\lvert Feedback_{it}\rvert$	0.0073^{***} (0.0025)
$Feedback^2_{it}$	$-1.12 \times 10^{-6**}$ (4.91×10^{-7})
$\mathbb{1}\{Comment_{i,t-1}\}$	0.2234^{***} (0.0371)
$Karma_{i,t-1}$	-5.5×10^{-5} (7.46×10^{-5})
$Karma^2_{i,t-1}$	5.67×10^{-11} (3.73×10^{-11})
$DistFromThreshold_{i,t-1}$	3.47×10^{-5} (7.59×10^{-5})
$DistFromThreshold^2_{i,t-1}$	-9.13×10^{-11} (1.45×10^{-10})
User fixed effects	✓
Date fixed effects	✓
Number of Digits fixed effects	✓
Observations	49,197
Pseudo R^2	0.49065

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All coefficients for the specification reported in Table 5. Standard errors are clustered at the user level.

Table C.5: Effect of Feedback on Where Users Post: All Coefficients

Dependent Variables:	Samesubmission	Samesubreddit	Diffsubreddit
Model:	(1)	(2)	(3)
$\mathbb{1}\{NegativeFeedback_{it}\}$	0.5680^{**}	1.472^{***}	2.296^{***}
	(0.2388)	(0.1018)	(0.0887)
$\mathbb{1}\{NegativeFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	-0.3198	-0.6384	1.473^{***}
	(1.132)	(0.4908)	(0.4099)
$\mathbb{1}\{PositiveFeedback_{it}\}$	0.4493^{***}	1.573^{***}	2.414^{***}
	(0.1085)	(0.0598)	(0.0507)
$\mathbb{1}\{PositiveFeedback_{it}\} \times \mathbb{1}\{LDChanged_{it}\}$	0.5709^{*}	0.0505	0.5539^{***}
	(0.3064)	(0.1015)	(0.0946)
$\lvert Feedback_{it}\rvert$	-0.0404^{***}	-0.0010^{***}	0.0025^{***}
	(0.0085)	(0.0002)	(0.0004)
$Feedback_{it}^2$	$9.3 \times 10^{-6***}$	$1.83 \times 10^{-7***}$	$-5.62 \times 10^{-7***}$
	(1.98×10^{-6})	(5.26×10^{-8})	(1.72×10^{-7})
$\mathbb{1}Comment_{i,t-1}$	3.115^{***}	0.3801^{***}	-0.3081^{***}
	(0.1584)	(0.0493)	(0.0502)
$Karma_{i,t-1}$	0.0002	$6.68 \times 10^{-5**}$	$-7.01 \times 10^{-5***}$
	(0.0002)	(2.94×10^{-5})	(2.37×10^{-5})
$Karma_{i,t-1}^2$	-1.34×10^{-9}	5.26×10^{-11}	1.09×10^{-11}
	(1.05×10^{-9})	(4.12×10^{-11})	(1.08×10^{-11})
$DistFromThreshold_{i,t-1}$	-0.0002	1.66×10^{-5}	2.21×10^{-5}
	(0.0004)	(3.93×10^{-5})	(3.91×10^{-5})
$DistFromThreshold_{i,t-1}^2$	5.53×10^{-9}	$-3.89 \times 10^{-10*}$	1.57×10^{-12}
	(1.23×10^{-8})	(2.2×10^{-10})	(6.63×10^{-11})
$\mathbb{1}CommentSameSubmission_{i,t-1}$	-1.404^{***}		
	(0.1974)		
$\mathbb{1}CommentSameSubreddit_{i,t-1}$		-0.0727	
		(0.0449)	
$\mathbb{1}CommentDiffSubreddit_{i,t-1}$			-0.0183
			(0.0449)
User fixed effects	✓	✓	✓
Date fixed effects	✓	✓	✓
Observations	17,709	34,004	49,687
Pseudo R^2	0.20665	0.25038	0.33359

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

All coefficients for the specifications reported in Table 6. Standard errors are clustered at the user level.

Table C.6: Effect of Feedback on Intensity: All Coefficients

Dependent Variables: Model:	|Sentiment| of Next Mention (1)	Magnitude of Next Mention (log) (2)
$\|s_c\|$	0.1208^{***} (0.0127)	
$\mathbb{1}\{NegativeFeedback\}$	0.0071 (0.0046)	0.0097^{**} (0.0047)
$\mathbb{1}\{PositiveFeedback\}$	0.0052 (0.0036)	0.0085^{**} (0.0042)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0700^{***} (0.0232)	
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0283 (0.0183)	
$\log(m_c)$		0.0839^{***} (0.0087)
$\mathbb{1}\{NegativeFeedback\} \times \log(m_c)$		-0.0502^{***} (0.0186)
$\mathbb{1}\{PositiveFeedback\} \times \log(m_c)$		-0.0186 (0.0162)
$Score_{t-1}$	3.41×10^{-5} (3.04×10^{-5})	2.55×10^{-5} (2.48×10^{-5})
$Score_{t-1}^2$	-8.14×10^{-9} (7.92×10^{-9})	-6.72×10^{-9} (6.73×10^{-9})
$\|Feedback_t\|$	-0.0008^{*} (0.0004)	-0.0008 (0.0005)
$Feedback_t^2$	1.75×10^{-6} (1.23×10^{-6})	$3.16 \times 10^{-6**}$ (1.53×10^{-6})
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Observations	31,707	31,707
R^2	0.25962	0.25101

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Column 1 reports all coefficients for the specification reported in Table 7. Column 2 is analogous to Column 1, however, unlike Column 1 which measures intensity using the absolute value of the sentiment score, this column uses the logged magnitude score. Standard errors are clustered at the user level.

Table C.7: Effect of Feedback on Sentiment: All Coefficients

	|Sentiment| of Next Mention	
	(1)	(2)
$\|s_c\|$	0.1207*** (0.0127)	0.1223*** (0.0128)
$\mathbb{1}\{NegativeFeedback\}$	0.0050 (0.0046)	0.0114* (0.0063)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0724** (0.0289)	0.0622** (0.0301)
$\mathbb{1}\{NegativeFeedback\} \times \|s_c\|$	-0.0609*** (0.0233)	-0.0706*** (0.0273)
$\mathbb{1}\{ScoreBecameNegative\} \times \|s_c\|$	-0.2652*** (0.0938)	-0.2592*** (0.0969)
$\mathbb{1}\{PositiveFeedback\}$	0.0054 (0.0037)	0.0047 (0.0043)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0050 (0.0170)	9.5×10^{-5} (0.0171)
$\mathbb{1}\{PositiveFeedback\} \times \|s_c\|$	-0.0276 (0.0183)	-0.0170 (0.0222)
$\mathbb{1}\{ScoreBecamePositive\} \times \|s_c\|$	-0.0637 (0.0979)	-0.1005 (0.1002)
$Score_{t-1}$	3.42×10^{-5} (3.05×10^{-5})	0.0001 (0.0006)
$Score^2_{t-1}$	-8.21×10^{-9} (7.94×10^{-9})	-2.42×10^{-5} (9.07×10^{-5})
$\|Feedback_t\|$	-0.0008* (0.0004)	-0.0003 (0.0007)
$Feedback^2_t$	1.78×10^{-6} (1.24×10^{-6})	3.6×10^{-7} (1.8×10^{-6})
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Sample	All	$\|Score_{t-1}\| \leq 10$
Observations	31,707	27,762
R^2	0.25996	0.27860

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Column 1 reports estimates for all the coefficients for the specifications reported in Table 8. Column 2 reports the coefficients for a subset, where comment's score at $t-1$ is less than or equal to 10. Standard errors are clustered at the user level.

Table C.8: Effect of Feedback on Magnitude: All Coefficients

	Magnitude of Next Mention (log)	
	(1)	(2)
$\mathbb{1}\{NegativeFeedback\}$	0.0084^{*}	0.0150^{**}
	(0.0047)	(0.0063)
$\mathbb{1}\{ScoreBecameNegative\}$	0.0567^{*}	0.0487
	(0.0293)	(0.0303)
$\mathbb{1}\{NegativeFeedback\} \times \log(m_c)$	-0.0458^{**}	-0.0607^{***}
	(0.0184)	(0.0191)
$\mathbb{1}\{ScoreBecameNegative\} \times \log(m_c)$	-0.1816^{*}	-0.1781^{*}
	(0.0980)	(0.0993)
$\mathbb{1}\{PositiveFeedback\}$	0.0089^{**}	0.0085^{*}
	(0.0043)	(0.0048)
$\mathbb{1}\{ScoreBecamePositive\}$	-0.0250	-0.0173
	(0.0275)	(0.0281)
$\mathbb{1}\{PositiveFeedback\} \times \log(m_c)$	-0.0198	-0.0122
	(0.0162)	(0.0208)
$\mathbb{1}\{ScoreBecamePositive\} \times \log(m_c)$	0.0967	0.0502
	(0.1671)	(0.1675)
$\log(m_c)$	0.0839^{***}	0.0844^{***}
	(0.0087)	(0.0091)
$Score_{t-1}$	2.59×10^{-5}	-0.0002
	(2.47×10^{-5})	(0.0006)
$Score_{t-1}^2$	-6.83×10^{-9}	2.21×10^{-5}
	(6.72×10^{-9})	(8.68×10^{-5})
$\lvert Feedback_t \rvert$	-0.0008	-0.0001
	(0.0005)	(0.0008)
$Feedback_t^2$	$3.2 \times 10^{-6**}$	1.74×10^{-6}
	(1.52×10^{-6})	(2.11×10^{-6})
User fixed effects	✓	✓
Subreddit fixed effects	✓	✓
Days After Posting fixed effects	✓	✓
Date fixed effects	✓	✓
Sample	All	$\lvert Score_{t-1} \rvert \leq 10$
Observations	31,707	27,762
R^2	0.25119	0.26966

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This table is analogous to Table C.7. However, unlike Table C.7 which measures intensity using the absolute value of the sentiment score, this table uses the logged magnitude score. Standard errors are clustered at the user level.