

# Sentiment Dynamics and Shifts across Instances on Mastodon

Seeun Kim<sup>1</sup>, Li Zeng<sup>1</sup>, Sijia Ma<sup>1</sup>, Giulia Sturlese<sup>2</sup>

<sup>1</sup>Tilburg University, Netherlands

<sup>2</sup>IUSS Pavia School of Advanced Studies, Italy

## Abstract

This study investigates how sentiment shifts unfold on Mastodon, a decentralized social network, during the 2023 Gaza conflict. We examine how the federated structure shapes the spread and transformation of sentiment in crisis-related discourse. We find that posts with broader reach tend to show more volatile sentiment, major conflict events coincide with sharp sentiment changes, and reply interactions often diverge in sentiment from the original post. These findings highlight how decentralized infrastructures mediate sentiment dynamics in online communications.

## Introduction

The dynamics of sentiment expression and evolution on social networks have been extensively studied in the context of centralized platforms like Twitter and Facebook (e.g., Thelwall, Buckley, and Paltoglou (2012); Kramer, Guillory, and Hancock (2014)). However, the rise of decentralized social networks, such as Mastodon, presents a unique opportunity to revisit these dynamics in decentralized environments. Unlike centralized platforms that operate under uniform content delivery algorithms and moderation policies, decentralized systems are composed of semi-independent instances, each with its own norms, user bases, and degrees of inter-instance connectivity (Zignani, Gaito, and Rossi 2018; La Cava, Greco, and Tagarelli 2021). This structural difference raises new questions about how sentiment diffuses and shifts over time across federated networks.

The 2023 Gaza conflict triggered a surge of emotionally charged discourse on Mastodon. While some posts (also known as “toots”) remained largely confined within local instances, others spread across instance boundaries through boosts, replies, and hashtags. This presents an ideal context to examine how sentiment propagates in a decentralized system, and whether cross-instance diffusion correlates with notable shifts in sentiment (Daou 2019; Knutson et al. 2024).

In this study, we analyze toots related to the Gaza conflict across a broad range of Mastodon instances to understand how sentiment evolves in a decentralized environment. Specifically, we examine (1) whether more widespread toots show more volatile sentiment shifts, (2) whether sentiment

peaks align with salient real-world events, and (3) whether reply dynamics amplify or alleviate the sentiment of the original message. These questions guide our analysis of the sentiment dynamics and shifts across instances on Mastodon.

## Related Work

Prior research has shown that the broader a message’s reach, the more emotionally diverse the responses it tends to attract (Ferrara and Yang 2015). Garimella et al. (2018) showed that exposure to diverse audiences leads to a greater sentiment divergence. In federated environments like Mastodon, instance-level structures enable such diversity in exposure, which raises the question of whether toots that travel more widely across instances also exhibit greater variability in sentiment over time (Zignani, Gaito, and Rossi 2018).

Another line of research connects virality to emotional intensity. Knutson et al. (2024) found that polarized content tends to produce high-arousal negative affects and is more likely to be reposted. Similarly, negative or emotionally charged headlines have been shown to increase news consumption (Robertson et al. 2023). Daou (2019) demonstrated that major real-world events can be detected by identifying strong emotional signals in social media. These findings suggest a feedback loop between high-impact events and online sentiment, motivating us to examine whether key conflict events align with sharper sentiment changes.

Finally, the structure of replies and conversation threads also plays a role in shaping sentiment trajectories. De Choudhury et al. (2016) showed that replies can reinforce or alleviate sentiment by enabling support and shared emotional framing. Chu et al. (2024) found that replies on social media tend to reflect or reinforce the sentiment of the original post. Backstrom et al. (2013) also found that the thread structures of conversations on social media can intensify engagement with the original message. These studies suggest that replies may shape sentiment in federated discourse as well.

## Methods

### Data

We collected Mastodon posts (“toots”) related to the 2023 Gaza conflict from nine instances: mastodon.social, masto.nyc, seattle.wa.us, sfa.social, gardenstate.social, better.boston, mastodon.nz, mastodon.au, and mastodon.ie.

These instances were selected for their high activity levels and geographic diversity. Due to the federated nature of Mastodon, we were able to capture toots not only from these instances but also from other connected ones, resulting in a broad and diverse dataset.

Using Python scripts and the Mastodon API, we implemented keyword-based queries to collect toots related to the Gaza conflict. The ten keywords used in data collection were *gaza*, *israel*, *palestine*, *hamas*, *genocide*, *freepalestine*, *cease-fire*, *idf*, *warcrimes*, and *palestinians*. These keywords were selected based on empirical analysis conducted during the pre-collection phase, as well as external sources, such as news media. The data collection yielded 200,096 toots posted between September 1 and December 19, 2023, containing terms related to the Gaza conflict. Each toot includes user-level, content-level, and thread-level metadata. After removing duplicates, the final dataset comprises 98,937 toots, covering 1,311 instances and 18,789 users. The dataset contains both original posts (seed toots, 56%) and replies (44%). Notably, 42,328 seed toots (42.78%) received no replies, suggesting high variability in engagement.

To analyze sentiment diffusion, we constructed reply trees for each thread with at least one reply. Each tree is rooted at the seed toot, and replies form successive child nodes. We defined the **Cross-Instance Spread** ( $C_i$ ) as the number of replies originating from instances other than the instance of the seed toot, capturing the extent of diffusion across the federated network.

Based on  $C_i$  values, we categorized threads into four spread groups. The **no spread** group ( $C_i = 0$ ) made up 58.27% of all threads. The remaining 41.73% of the threads were divided into three groups: **low spread** ( $C_i \in [1, 2]$ ), representing 45.93% of the remaining threads; **medium spread** ( $C_i \in [3, 6]$ ), comprising 23.14%; and **high spread** ( $C_i > 6$ ), comprising 27.83%. These groups reflected increasing levels of cross-instance diffusion and were used consistently in our analyses of sentiment dynamics.

## Sentiment Labeling

To analyze the sentiment of the toots, we compared three approaches: VADER, mBERT, and XLM-T.

- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based tool designed for short, social media-style text (Hutto and Gilbert 2014).
- mBERT (Multilingual Bidirectional Encoder Representations from Transformers) is a powerful language model that captures contextual meaning in text, widely used in sentiment classification tasks (Devlin et al. 2019).
- XLM-T (Cross-Lingual Language Model for Twitter) is pretrained on multilingual Twitter data and performs well on informal, short-form, and cross-lingual social media content (Barbieri, Espinosa Anke, and Camacho-Collados 2022).

To determine the most reliable method, two independent evaluators manually labeled a random sample of 200 toots for sentiment (positive, negative, or neutral). In cases of disagreement, the third evaluator adjudicated the final label. The overall inter-coder agreement was substantial, with a Fleiss'

Table 1: Comparison of sentiment classification models on the manual validation set.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F <sub>1</sub>
XLM-T	0.783	0.801	0.783	0.761
mBERT	0.614	0.510	0.614	0.535
VADER	0.529	0.526	0.529	0.527

kappa score of 0.80. The performance of the three candidate models, XLM-T, mBERT, and VADER, on this manually annotated validation set is presented in Table 1. Among them, the XLM-T model demonstrated the highest consistency with the human annotations across all measures. Therefore, we selected XLM-T for sentiment classification in our study. Furthermore, XLM-T holds promise for even greater performance with future fine-tuning (Barbieri, Espinosa Anke, and Camacho-Collados 2022).

## Sentiment Volatility Analysis

To assess how sentiment fluctuated across spread groups over time, we defined **sentiment volatility** as the absolute change in average sentiment from one week to the previous week. Specifically, for each spread group  $g$  and week  $w$ :

$$\Delta S_g(w) = |S_g(w) - S_g(w - 1)|$$

where  $S_g(w)$  is the average sentiment score for group  $g$  in week  $w$ . Weeks were defined as fixed 7-day intervals from September 30 to mid-December 2023 (e.g., Week 1: Sept. 30 – Oct. 6; Week 2: Oct. 7 – Oct. 13).

We used the resulting set of  $\Delta S_g(w)$  values as a measure of the weekly sentiment volatility for each spread group.

To summarize volatility within groups, we computed:

- **Mean delta** ( $\bar{\Delta S}_g$ ): average week-to-week sentiment change per group.
- **Standard deviation of deltas** ( $\sigma_{\Delta S}_g$ ): variability in week-to-week sentiment changes.

To statistically compare sentiment volatility across spread groups, we applied the non-parametric Kruskal-Wallis test on weekly changes in average sentiment. For pairwise comparisons, we used Dunn’s post hoc test with Bonferroni correction to control for multiple testing. This analysis evaluates whether groups with broader diffusion show significantly greater sentiment volatility.

## Event-based Sentiment Analysis

To examine how key conflict events influenced sentiment, we identified three major events based on international news reports:

1. **October 7, 2023**: The conflict broke out.
2. **November 24, 2023**: The ceasefire began.
3. **December 3, 2023**: The conflict resumed.

For each event, we defined a 3-day pre-event window ( $t-3$  to  $t-1$ ) and a 3-day post-event window ( $t+1$  to  $t+3$ ), where  $t$  denotes the event day. This shorter, non-overlapping window avoids contamination between events and focuses on capturing sharp shifts in sentiment immediately surrounding

each moment. The event day itself ( $t$ ) was excluded to reduce bias from unusually high activity or atypical sentiment distributions on that specific day. In addition, sentiment expressed on the event day may reflect a mix of anticipatory, real-time, and delayed reactions, which can obscure before-and-after comparisons.

To test whether these events were associated with significant changes in sentiment, we conducted a Mann–Whitney U test comparing sentiment scores of posts made in the pre- and post-event windows. This non-parametric test is appropriate given that the two samples (pre- and post-event posts) are independent and not paired. We performed this test separately for each event. This allowed us to assess whether the key events triggered significant sentiment shifts.

### Reply Sentiment Shift Analysis

To examine how the sentiment shifts in replies relate to the sentiment of the original post, we analyzed threads containing at least two replies. For each thread, we calculated the **sentiment shift** as the difference between the sentiment of the seed toot and the average sentiment of its replies:

$$\Delta S_{\text{reply-seed}} = \bar{S}_{\text{replies}} - S_{\text{seed}}$$

This  $\Delta S_{\text{reply-seed}}$  value represents the direction and magnitude of the sentiment change between an original post and the collective sentiment of its responses. We also computed the standard deviation of reply sentiment within each thread to assess intra-thread sentiment divergence.

To test whether the sentiment of replies systematically differs from that of the original posts, we applied the Wilcoxon signed-rank test on the aggregated set of sentiment differences between replies and original post across all threads. This non-parametric test is appropriate for paired data, as each  $\Delta S_{\text{reply-seed}}$  value is computed from a matched pair.

To evaluate whether the degree of sentiment shift varied across levels of cross-instance content spread, we conducted a Kruskal–Wallis test on the  $\Delta S_{\text{reply-seed}}$  distributions corresponding to each cross-instance spread category. When the result was significant, we followed up with Dunn’s post-hoc test with Bonferroni correction to identify group differences.

## Results

### Sentiment Volatility Analysis

To understand how sentiment evolves and varies by cross-instance spread, we analyzed the week-to-week volatility of sentiment scores. Figure 1 shows the distribution of sentiment volatility across different levels of cross-instance spread. The median and interquartile range of sentiment change increase progressively from the no spread group to the high spread group, indicating that greater content diffusion is associated with larger fluctuations in sentiment. This pattern is supported by the descriptive statistics: The average sentiment change ( $\bar{\Delta S}_g$ ) is highest in the high spread group (0.305), followed by medium (0.270), low (0.260), and no spread (0.219). Similarly, the standard deviation of sentiment change, which reflects the stability of the sentiment fluctuations over time, was largest for the high spread group (0.299) and smallest

Table 2: Post-hoc Dunn’s test (Bonferroni-corrected) for week-to-week sentiment volatility across spread groups.

	High	Medium	Low	No Spread
High	1.000	1.000	0.308	0.011
Medium	1.000	1.000	1.000	0.315
Low	0.308	1.000	1.000	1.000
No Spread	0.011	0.315	1.000	1.000

for the no spread group (0.199). Together, these findings suggest that higher levels of content spread across instances are associated with greater sentiment variability and instability.

To test whether sentiment volatility differed significantly across spread groups, we first conducted a Kruskal–Wallis test. The result was statistically significant ( $H(3) = 10.38, p = 0.016$ ), indicating that at least one group significantly differs from the other groups in terms of sentiment volatility. We then performed a post-hoc Dunn’s test with Bonferroni correction to identify which groups differed. The result in Table 2 showed that high spread and no spread groups has a significant difference ( $p = 0.011$ ). This suggests that the sentiment of widely spread content fluctuates more compared to confined content, while differences between adjacent spread levels (e.g., medium vs. low) were not large enough to be statistically significant, especially after Bonferroni correction.

In summary, these results provide evidence that greater content spread is associated with greater sentiment volatility. Content that diffuses across more instances tends to trigger larger variations in sentiment from week to week.

### Event-based Sentiment Analysis

Figure 2 shows the weekly average sentiment scores across the four spread groups throughout the observation period. Vertical dashed lines mark the timing of the three key conflict events: the outbreak of the war, the ceasefire, and the resumption of fighting. We observe that clear changes in sentiment responses often occur with a one- to two-day delay relative to the official event dates. For example, although the conflict

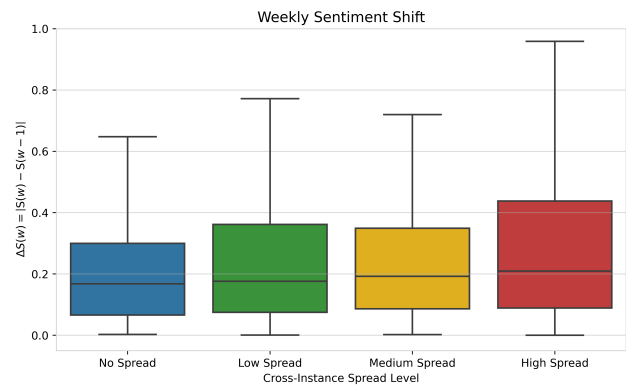


Figure 1: Distribution of sentiment volatility across spread groups.

began on October 7 and the fighting resumed on December 3, sharp sentiment changes appear on October 9 and December 4, respectively. In general, all spread groups expressed negative sentiment throughout the period, but clear differences emerged in how sentiment shifted around major events.

To test the statistical significance of these shifts, we conducted Mann-Whitney U tests comparing sentiment scores in three-day windows before and after each event. The outbreak of conflict on October 7 was followed by a statistically change in sentiment ( $U = 330075.0, p < 0.001$ ), as was the ceasefire declared on November 24 ( $U = 32771.0, p < 0.001$ ). In contrast, no significant shift was detected following the resumption of fighting on December 3 ( $U = 363.0, p = 0.824$ ), and the average change in sentiment during that period was relatively small (0.072).

To summarize, the event-based analysis indicates that major developments in the conflict, particularly the initial outbreak and the announcement of a ceasefire, were accompanied by significant shifts in sentiment.

### Reply Sentiment Shift Analysis

Lastly, we examined whether replies exhibited systematic sentiment differences from the original posts they responded to. Figure 3 shows that reply sentiment shift ( $\Delta S_{\text{reply-seed}}$ ) increases with spread level, with the highest median shift observed in the high spread group and the lowest in the no spread group. To complement this, we also calculated the average sentiment divergence within threads, which was the highest in the high spread group (0.38), followed by medium (0.36), no (0.34), and the lowest in the low spread group (0.26). This pattern indicates that greater cross-instance diffusion is associated not only with stronger sentiment shifts, but also with more heterogeneous sentiment responses.

To support the exploratory finding above, a Wilcoxon signed-rank test confirmed a significant shift in sentiment across all threads (Wilcoxon statistic = 25,635,912,  $p < 0.001$ ), indicating that sentiment often changes in replies

relative to the original message.

To test whether the sentiment shift differed significantly by spread groups, we conducted a Kruskal–Wallis test on  $\Delta S_{\text{reply-seed}}$  values. The result in Table 3 was statistically significant ( $H = 22.97, p < 0.001$ ), and post-hoc Dunn’s tests (Bonferroni corrected) revealed that the low spread group differed significantly from the no spread ( $p = 0.001$ ), medium spread ( $p = 0.010$ ) and high-spread ( $p = 0.021$ ) groups.

Together, these findings suggest that low-spread threads show the most constrained sentiment dynamics, with significantly smaller shifts compared to all other groups. This points to a potential threshold pattern, where limited diffusion corresponds to more emotionally uniform replies, while both no-spread and widely spread posts tend to elicit more varied sentiment responses.

Table 3: Post-hoc Dunn’s test (Bonferroni-corrected) for reply sentiment shift ( $\Delta S_{\text{reply-seed}}$ ) across spread groups.

	High	Medium	Low	No Spread
High	1.000	1.000	0.021	1.000
Medium	1.000	1.000	0.010	1.000
Low	0.021	0.010	1.000	0.001
No Spread	1.000	1.000	0.001	1.000

## Discussion

Our findings point to a strong association between cross-instance diffusion and sentiment dynamics on federated platforms. Posts that spread more widely across instances tend to present greater variation in sentiment, both over time and across replies. Major conflict-related events, such as the outbreak of war and the ceasefire, coincided with significant shifts in overall sentiment. These results suggest that diffusion is not only a matter of reach but may actively reshape how content is perceived and interpreted as it circulates.

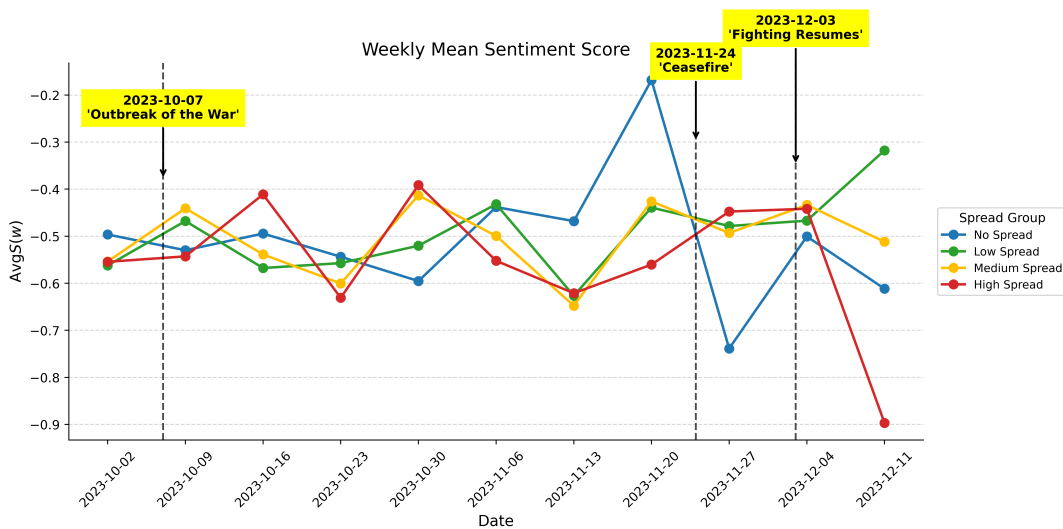


Figure 2: Weekly average sentiment score by cross-instance spread group with the three key events annotated.

The positive relationship between spread level and sentiment volatility implies that content reaching broader audiences is more likely to provoke heterogeneous emotional responses. This may reflect exposure to more ideologically or culturally diverse users across instances, or the contextual reinterpretation of content in different local settings. This aligns with prior work suggesting that exposure to heterogeneous networks leads to greater variation in public responses (Barberá et al. 2015).

We also observe that sentiment shifts in replies vary by spread level, with significantly smaller shifts in low-spread conversations compared to medium- and high-spread ones. This indicates that cross-instance diffusion may increase sentiment heterogeneity within conversations, potentially fostering sentiment complexity rather than simple alignment or polarization.

This study suggests several opportunities for future research. First, future research could include other types of events to analyze whether patterns of sentiment shift and divergence vary by context. While our dataset includes a wide range of instances, it may underrepresent smaller or less-connected communities. In terms of sentiment labeling, while we used a multilingual sentiment model, future work could improve accuracy by fine-tuning models or incorporating large language models to better capture tone, sarcasm, and culturally embedded expressions. Finally, future research could examine how instance-level characteristics, such as moderation practices or community norms, affect the likelihood of hosting diverging sentiment threads, shedding light on how federation shapes sentiment spread and variability.

## References

Backstrom, L.; Kleinberg, J.; Lee, L.; and Danescu-Niculescu-Mizil, C. 2013. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 13–22.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political

communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Barbieri, F.; Espinosa Anke, L.; and Camacho-Collados, J. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proc. of the 13 Language Resources and Evaluation Conference*, 258–266. Marseille, France: European Language Resources Association.

Chu, M.; Song, W.; Zhao, Z.; Chen, T.; and Chiang, Y.-c. 2024. Emotional contagion on social media and the simulation of intervention strategies after a disaster event: a modeling study. *Humanities and Social Sciences Communications*, 11(1): 968.

Daou, H. 2019. Detection of Sentiment Provoking Events in Social Media. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2503–2512. Hawaii International Conference on System Sciences.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ferrara, E.; and Yang, Z. 2015. Measuring emotional contagion in social media. *PloS one*, 10(11): e0142390.

Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proc. of the 2018 world wide web conference*, 913–922.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Knutson, B.; Hsu, T. W.; Ko, M.; and Tsai, J. L. 2024. News source bias and sentiment on social media. *PloS one*, 19(10).

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

La Cava, L.; Greco, S.; and Tagarelli, A. 2021. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied network science*, 6: 1–35.

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

Thelwall, M.; Buckley, K.; and Paltoglou, G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Info. Science and Technology*, 63(1): 163–173.

Zignani, M.; Gaito, S.; and Rossi, G. P. 2018. Follow the “mastodon”: Structure and evolution of a decentralized online social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 541–550.

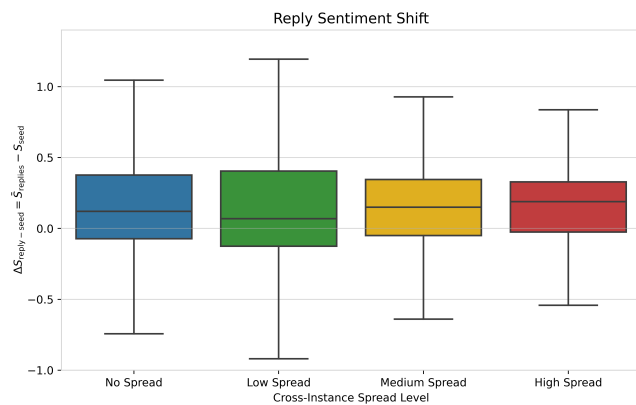


Figure 3: Distribution of reply sentiment shift ( $\Delta S_{\text{reply-seed}}$ ) across spread groups.