Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

You See It, They Don't: An Exploratory Study of User-to-User Variation in Instagram Comments

Brahmani Nutakki[1], **Manon Lilott Kempermann**[1], **Ingmar Weber**[1]
[1]Saarland University, Germany
bnutakki@cs.uni-saarland.de, make00009@stud.uni-saarland.de, iweber@cs.uni-saarland.de

Abstract

In March 2025, Meta announced a new AI system to rank the order of the comments shown to Instagram users. With existing research showing how feed personalization systems can lead to increased polarization, the introduction of this new system raises similar questions. This paper presents a small-scale exploratory study examining whether the ranking system produces systematic differences in visible comments shown to different users, particularly for news-related content. Using four sock-puppet accounts varying in gender and political leaning, we collect visible comments on posts from ten news and ten non-news accounts. This collection is repeated twice from two VPN locations to assess location effects. We ask 1) how many visible comments vary across different users, 2) is this variation higher for news accounts than non-news accounts, and 3) can user-attributes like gender, political leaning, and location systematically explain the observed variation. Contrary to our expectations, we find that visible comments on news posts are less likely to vary across users than those on non-news posts. Variation is better explained by account metrics like comment and follower counts than by user attributes. These findings provide an initial glimpse into personalized comment ranking on Instagram and motivate larger, more systematic audits of how comment personalization may shape online discourse. To support further research, we provide the code to collect comments and the data upon request.

1 Introduction

Over time, social media platforms have come to function as public squares, where users can engage in discussions on diverse topics. One such extremely popular platform is Instagram, with reports finding that the app is installed on almost 80% of smartphones worldwide, amounting to 3 billion Monthly Active Users (DataReportal 2025). As early as 2016, Instagram revealed that an average of 95 million photos (and videos) were shared per day, a figure that has likely grown since (Reuters 2016). Yet, surveys indicate that only 20% of users actively post on the platform (Inc 2023), implying that most users interact primarily by liking, sharing, or commenting on existing posts.

Of the above metrics, comments emerge as the epicenter of dialogue and engagement on the platform: even posts from accounts with 1000-2000 followers receive at least 1-3 comments on average (SocialInsider 2025). This ratio grows drastically, particularly with larger accounts, which can garner tens of thousands of comments. Previous studies have shown that attention-grabbing comments that are highly engaging are frequently negative and emotionally charged (Risch and Krestel 2020; Heraki and Zaghouani 2025). These comments can reduce a post's credibility (Naab et al. 2020), and a user's intent to share it (Boot, Dijkstra, and Zwaan 2021). Conversely, user-initiated corrections can also effectively counter misinformation if they include reliable sources (Seo et al. 2022).

In March 2025, Meta announced that an AI system was being used to personalize comments shown to Instagram users (Meta 2025). This system takes different input signals, such as the likelihood of a user to report, delete, reply, click on, or scroll past a comment, to decide how they should be ranked (Meta 2025). As of writing this paper, Instagram lets users choose among three settings for the comments: 'For You', 'Most recent', and 'Meta verified', with 'For You' being the default selection. Given existing literature on how social media's algorithmic personalization may contribute to increased polarization (Pournaki, Gaisbauer, and Olbrich 2025; Cinus et al. 2022), this announcement has raised similar questions. While there are works that question whether personalization is the primary driver for the observed increase in polarization (Garimella and Weber 2017), and the long-standing debate on the impact of personalization (Conover et al. 2021; Dahlgren 2021), recent intervention experiments identified causal links between them (Piccardi et al. 2025). Exposure to diverse political content can even trigger "backfire effects" increasing polarization (Bail et al. 2018). Prior research also indicates that comments appearing at the top can shape the narrative that follows (Naab et al. 2020; Zhang et al. 2018). Yet, compared with the extensive literature on feed ranking, there appears to be relatively limited research on comment ranking.

In this context, we wanted to understand how this new ranking system might influence narratives and potentially amplify polarizing perspectives. Given that this is a relatively new development and conducting large-scale sock-puppet audits is challenging, we begin with a small-scale exploratory study. For this study, our goal is to examine how visible comments vary across users viewing news and

non-news content. Nearly 55.8% of Instagram users use the app to stay up-to-date with news (DataReportal 2025), and news is often polarizing. Social media news comment sections have also been characterized as real-time barometers of public opinion (Hossain et al. 2024), with research showing that users who are typically highly engaged with news are more likely to actively comment on the posts (Kalogeropoulos et al. 2017). Given this, news acts as a natural starting point to assess whether user attributes, such as gender, political leaning, and location, affect the comments they see. To provide a baseline, we compare this against non-news content. Our research questions include **1)** To what extent do comments vary across users viewing the same post? **2)** Is user-user variation greater for news-related posts than non-news posts? and **3)** Are user attributes associated with systematic differences in the comments shown?

To do this, we analyze visible comments displayed to four sock-puppet users with different political leanings and gender profiles: Female Democrat, Female Republican, Male Democrat, and Male Republican. For each, we collect the visible comments (visible without scrolling, referred to as comments from now on) shown on the same posts from ten news accounts and ten non-news accounts. To test location effects, we repeat the collection from two VPN endpoints in New York State and Texas. We selected these two diverse US locations as most of the news accounts are US-based.

Contrary to our initial hypothesis, our results show that, on average, only 12% of the comments differ across users. Comments on news posts are actually less likely to vary than those on non-news posts. Account metrics such as follower and comment counts are highly associated with variation but in opposite directions. By comparison, user attributes have only modest effects and do not appear to be the primary predictors.

To our knowledge, this is the first study to investigate personalized comment ranking on Instagram in this context. We hope this exploratory study highlights the need for more systematic research on this development. As platforms increasingly introduce personalization to boost engagement, understanding how these decisions shape discourse and their impact on an already fragmented society becomes vital. Both the code used to collect comments and the data itself are available upon request.

2 Data Collection

The data collection pipeline consists of three main steps: (1) creating sock-puppet user accounts, (2) collecting recent posts from selected Instagram accounts, and (3) collecting the top visible comments for each post as viewed by the sock-puppet accounts. The steps are described in detail in the subsections below.

Sock-Puppet Accounts Creation

To understand how comments might differ across different users, we created four new sock-puppet accounts using different email addresses. To maintain uniformity across accounts, we used the same date of birth and used gibberish usernames to avoid any username-specific inference by the platform. These accounts were set up with four personalities: Female Democrat, Female Republican, Male Democrat, and Male Republican. The bio was left empty, and the gender of the account was assigned in the settings, matching the gender assigned to the email address. For political-leaning preferences, we manually followed politicians from the Democratic and Republican parties for the algorithm to infer the account's preference. No other actions such as liking, sharing or commenting were performed to avoid unforeseen effects. Despite this limited interaction, the user and suggested feeds of the sock-puppet accounts showed personalization, with Republican accounts having predominantly Republican posts and the Democratic accounts following a similar pattern. The list of politicians followed can be seen in the appendix.

Posts Collection

We selected ten Instagram accounts belonging to news outlets across the political spectrum, using bias ratings from Media Bias/Fact Check (Check 2026). To compare these with non-news content, ten non-news accounts from different niches, such as sports, entertainment, food, and pets, were also selected (for both see Table 1). We ensured that the follower counts of news and non-news accounts collected roughly match to reduce unintended variability.

After identifying the accounts, we collected their ten most recent posts that were posted at least 24 hours before the point of data collection to ensure that their comment counts are saturated. While collecting posts that were not yet saturated could have provided more insights, we refrained from doing so to minimize uncertainty. More details about saturation can be found in Section 4. In total, this resulted in 200 posts from twenty different news and non-news accounts.

Comments Collection

Once the sock-puppet accounts and posts were available, we built a Selenium crawler to open each post and collect the caption, timestamp, and the number of comments. After logging in with each sock-puppet account, the crawler also collected the comments that were visible on each post without scrolling, along with their timestamps and the username of the commenter. We ran this collection process for every post under each sock-puppet account using two proxy locations: one in New York and one in Texas. The comments were collected at least 24 hours post the creation of the accounts to allow for personalization. This produced eight separate crawls in total—four accounts multiplied by two locations—each collecting top visible comments (usually 10-15) from the same set of 200 posts.

For the analysis, we limited attention to the top ten visible comments, since most posts contained at least ten. We collected only text-based comments, as the crawler could not reliably capture GIFs or hidden comments. Manual checks indicated that only about 30% of posts contained GIFs in the top comments; among those, most had only one or two. We also removed generic automated comments such as "You can review or change your choices at any time in your cookie settings." After filtering, we obtained around 930 comments

News Account	MBFC Bias	#followers	Non-News Account	Account Type	#followers
MSNBC	Left	2.4M	Peacock	Entertainment	2.4M
Huffington Post	Left	3.3M	Nytcooking	Food	4.6M
CNN	Center-Left	21.7M	Espn	Sports	28.4M
Washington Post	Center-Left	7.3M	Catloversclub	Pets	8M
Forbes	Center	7.3M	Thedogist	Pets	7.6M
The Hill	Center	304K	Thegradecricketer	Sports	323K
Washington Times	Center-Right	120K	Pbsfood	Food	156K
New York Post	Center-Right	2M	Hulu	Entertainment	2.6M
Fox News	Right	10.8M	Ladbible	Entertainment	15.2M
Breitbart	Right	1.8M	Accesshollywood	Entertainment	1.8M

Table 1: This table provides the accounts used in the data collection process, along with other details. Follower counts are reported as of January 4, 2025.

from news accounts and 970 comments from non-news accounts for each crawl. After combining comments across all crawls, we obtained 980 and 1037 unique comments from news and non-news accounts, respectively.

3 Empirical Analysis and Results

This section describes the approaches used and the insights they provide on the visibility of comments across different posts and different users.

Descriptive Analysis

As mentioned in Section 2, we collect comments from eight user crawls. To answer RQ1, we start by calculating the proportion of variation between each pair of crawls. For two crawls A and B, and for each post p, let $C_A(p)$ be the set of top comments on p collected in crawl A and $C_B(p)$ for those collected in crawl B. We define the variation between the two crawls on post p, denoted by $V(p)$, as the symmetric difference between the two sets (comments that appear in exactly one crawl):

$$V(p) = (C_A(p) \cup C_B(p)) - (C_A(p) \cap C_B(p))$$

We then normalize by the total number of collected comments across the two crawls for the post $T(p) = |C_A(p)| + |C_B(p)|$ to obtain the per-post variation proportion as $V(p)/T(p)$. The overall average variation proportion between crawls A and B is then computed by averaging across all posts:

$$\frac{1}{N} \sum_{p=1}^{N} V(p)/T(p)$$

, where N is the total number of posts across all accounts.

We repeat this for each pair of crawls to characterize how variation proportions differ across each pair. The resulting heatmaps for news and non-news accounts are shown in Figure 1. Overall, on average, only 12% of comments vary between crawls. Comments on non-news accounts exhibit higher average variation than news accounts across crawls, suggesting that post type may be an important factor. While the heatmaps reveal some variation between crawls, the patterns are subtle and do not show strong structural variation.

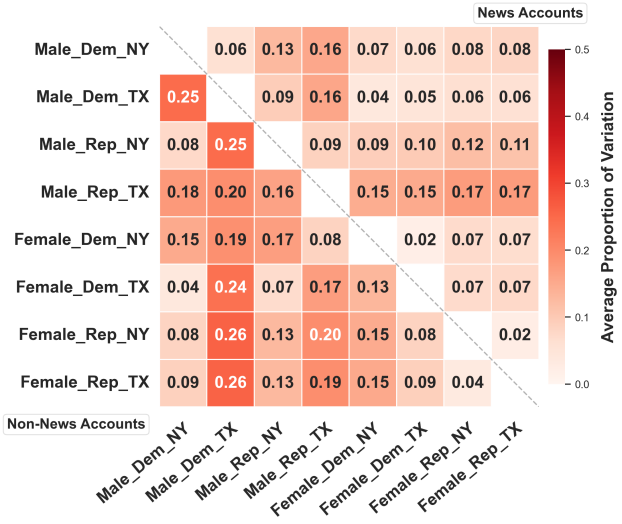

Figure 1: This heatmap shows the average variation proportion of comments across all posts between each pair of crawls. The upper triangle contains values for posts from News Accounts, and the lower triangle for those from Non-News Accounts.

Post-level Regression Analysis

To understand variations at a post level and answer RQ2 and RQ3, we fit a Bayesian generalized linear mixed model with a beta-binomial likelihood to model the variation observed for each post across a given crawl-pair comparison. Beta-binomial is chosen rather than binomial to account for over-dispersion. Each observation corresponds to one post and one crawl pair comparison; the outcome is variation $V(p)$ out of total trials $T(p)$ for post p, for all 28 crawl pairs (C_2^8). Fixed effects include non-directional labels for Location, Gender, and Leaning, derived from the two crawls being compared (e.g., comparing 'Female_Dem_NY' with 'Male_Rep_TX' produces the labels 'Female_Male', 'Dem_Rep,' and 'NY_TX'), along with comment and follower counts (log-transformed and standardized), and whether the post originated from a news or non-news account. To account for repeated observations

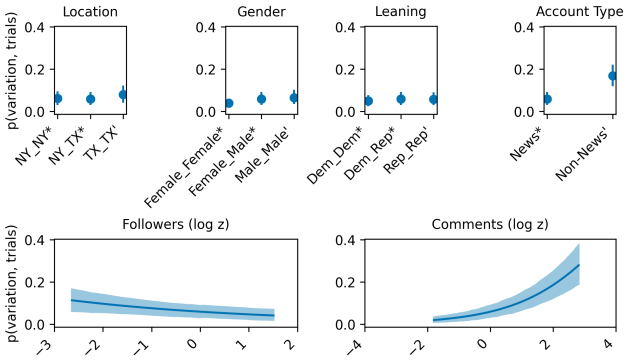

Figure 2: Posterior mean predictions (95% HDI) from the beta-binomial regression across categorical and continuous covariates, holding other predictors constant. The categories with ' are the reference categories, and * indicates categories whose credible interval for the odds ratio excludes 1.

from the same post, we added a random intercept for post ID. Model parameters and convergence statistics are included in the appendix.

The posterior mean prediction plot (see Figure 2) shows that account type and engagement metrics are more strongly associated with the predicted variation proportion. Holding other covariates constant, the model's posterior mean variation proportion is higher for non-news posts (0.17) than for news posts (0.06). It also increases steadily as the comment count increases. In contrast, follower counts show a negative association. Location, Gender, and Political Leaning effects are comparatively small in absolute terms (on the order of only a few percentage points on the probability scale). For several categories of Location, Gender, and Political Leaning, the 95% highest density intervals (HDI) for the odds ratio exclude 1, indicating evidence of association. However, the estimated magnitudes are small, suggesting limited practical significance, similar to the insights from the heatmaps. A forest plot of posterior odds ratios is provided in the appendix.

Comment-level Regression Analysis

To understand if the nature of comments is associated with variation, we ran additional analysis using posts from both news and non-news accounts. Each comment is labeled as 'Supportive', 'Against', or 'Neutral' relative to the post caption using OpenAI API (Exact prompt can be found in the appendix). These annotations were validated by comparing them against 30 manually annotated comments, and no discrepancies were found. For each crawl pair, we tracked each unique comment and coded whether it appeared in both crawls. The outcome was a binary indicator for non-overlap, where 0 meant present in both crawls and 1 meant present in only one crawl, so the model estimates the probability that a comment is missing from one of the two crawls. We fit a Bayesian Bernoulli (logistic) model with the same sampling configuration as before. Fixed effects are the same as the above, along with the comment label.

The results show that supportive comments have a lower estimated probability of non-overlap than against (and neutral) comments, indicating they are more consistently present across crawls. The effects of the remaining variables mirror the previous findings. The corresponding probability plot is provided in the appendix, due to space constraints.

4 Discussion and Conclusion

To the best of our knowledge, this is the first systematic study to investigate whether different users see different comments on Instagram posts following the introduction of the AI-driven comment ranking system.

Looking at the research questions outlined above, we find that overall variation is low: on average, 12% of the comments differ between crawls (RQ1). Comments on news accounts exhibit less variation than those on non-news accounts (RQ2). Account metrics such as follower and comment counts are also associated with variation, but in opposite directions: higher comment counts indicate higher variation, while lower follower counts indicate higher variation. One possible interpretation is an exploit-explore trade-off (Gisselbrecht et al. 2015): posts with many comments provide richer engagement signals, which may prompt the system to exploit these signals to personalize more. Whereas, smaller accounts might need more exploration and thus higher observed variation. While we do observe limited differences associated with location, gender, and political leaning (RQ3), given the current design, we can not establish their robustness.

To mitigate the effects of the rapidly evolving comments section, we only collect posts that are older than 24 hours. We chose this threshold by estimating the time needed to reach 95% of comment count after 72 hours of posting (approximately 12–17 hours on average), by collecting comments from newly published posts and tracking their growth in five-hour intervals. Similarly, the data collection for the 200 posts in the dataset for all eight crawls was done with a short time window of 4-5 hours to avoid the addition of new comments or other unknown factors. A brief check showed that the number of comments that were added during these 4-5 hours was extremely low (less than 10 across all posts).

Various factors may be associated with the observed variations. Since our sock-puppet accounts were newly created, the observed limited variance could also be attributed to the algorithm's limited knowledge about the users. We only use four sock-puppet accounts, which also limits the robustness of our results. We also observed that the commenters mostly aligned with the political bias of the news account (e.g, predominantly right-leaning comments on posts from Breitbart), which also explains supportive comments having a low probability of being present in only one crawl. Choosing a more contentious domain where both pro- and anti-comments are prevalent would have provided more interesting results. Lastly, in this study, we only look at the presence of a comment, i.e, whether the comment appears across crawls, not ranking order. A preliminary analysis of common comments across crawls showed that ranking variation is substantial. The average variance proportion (considering difference in rank as variation) was 0.5 for news accounts

and 0.7 for non-news accounts, indicating that the order of visibility differs largely across users.

Despite comments playing an important role in driving engagement and shaping narratives, there is surprisingly little work examining how comment-ranking systems influence what users see. Our work tries to explore the impact of the new comment-ranking system of Instagram by setting up a small-scale sock-puppet study. While the findings are modest, this study provides a starting framework for assessing how comment ranking could reinforce or amplify existing biases, and we hope that this study underscores the need for more such systematic research.

Acknowledgments

Ingmar Weber, Brahmani Nutakki, and Manon Lilott Kempermann are supported by funding from the Alexander von Humboldt Foundation and its founder, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung).

References

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Boot, A. B.; Dijkstra, K.; and Zwaan, R. A. 2021. The processing and evaluation of news content on social media is influenced by peer-user commentary. *Humanities and Social Sciences Communications*, 8(1): 209.

Check, M. B. 2026. https://mediabiasfactcheck.com/.

Cinus, F.; Minici, M.; Monti, C.; and Bonchi, F. 2022. The Effect of People Recommenders on Echo Chambers and Polarization. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 90–101.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini, A. 2021. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 89–96.

Dahlgren, P. M. 2021. A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, 42(1): 15–33.

DataReportal. 2025. Digital 2026: Global Overview Report. https://datareportal.com/reports/digital-2026-global-overview-report.

Garimella, V. R. K.; and Weber, I. 2017. A Long-Term Analysis of Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 528–531.

Gisselbrecht, T.; Denoyer, L.; Gallinari, P.; and Lamprier, S. 2015. WhichStreams: A Dynamic Approach for Focused Data Capture from Large Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1): 130–139.

Heraki, H. A.; and Zaghouani, W. 2025. Analyzing Digital Polarization on Hijab: A Dataset of Annotated YouTube Comments. *Proceedings of the International AAAI Conference on Web and Social Media*, 19: 2350–2360.

Hossain, I.; Puppala, S.; Alam, M. J.; Talukder, S.; and Talukder, Z. 2024. A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 596–609.

Inc, G. 2023. Social Media Users More Inclined to Browse Than Post Content. https://news.gallup.com/poll/467792/social-media-users-inclined-browse-post-content.aspx.

Kalogeropoulos, A.; Negredo, S.; Picone, I.; and Nielsen, R. K. 2017. Who Shares and Comments on News?: A Cross-National Comparative Analysis of Online and Social Media Participation. *Social Media + Society*, 3(4): 2056305117735754.

Meta. 2025. Instagram Comments. https://transparency.meta.com/features/explaining-ranking/ig-comments/.

Naab, T. K.; Heinbach, D.; Ziegele, M.; and Grasberger, M.-T. 2020. Comments and Credibility: How Critical User Comments Decrease Perceived News Article Credibility. *Journalism Studies*, 21(6): 783–801.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J.; Tsai, J. L.; and Bernstein, M. S. 2025. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776): eadu5584.

Pournaki, A.; Gaisbauer, F.; and Olbrich, E. 2025. How Influencers and Multipliers Drive Polarization and Issue Alignment on Twitter/X. *Proceedings of the International AAAI Conference on Web and Social Media*, 19: 1599–1615.

Reuters. 2016. Instagram's User Base Grows to More than 500 Million. *Reuters*.

Risch, J.; and Krestel, R. 2020. Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 579–589.

Seo, H.; Xiong, A.; Lee, S.; and Lee, D. 2022. If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 896–907.

SocialInsider. 2025. https://www.socialinsider.io/social-media-benchmarks/instagram.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. J. 2018. Characterizing Online Public Discussions through Patterns of Participant Interactions. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW): 198:1–198:27.

Ethics Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in the Discussion Section.

(e) Did you describe the limitations of your work? Yes, in the Discussion Section.

(f) Did you discuss any potential negative societal impacts of your work? NA

(g) Did you discuss any potential misuse of your work? NA

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes

(b) Have you provided justifications for all theoretical results? Yes

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

(e) Did you address potential biases or limitations in your theoretical framework? Yes, in the Discussion Section.

(f) Have you related your theoretical results to the existing literature in social science? Yes

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, code and data will be provided upon request.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in the appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? The model does not require large computation.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Appendix

List of Politicians Followed by Sock-Puppet Accounts

For the sock-puppet account to have a Democratic-leaning, we followed politicians including Former President Joe Biden (@joebiden), Former Vice President Kamala Harris (@kamalaharris), Former President Barack Obama (@barackobama), Representative Alexandria Ocasio-Cortez (@aoc), Senator Elizabeth Warren (@senwarren), Senator Bernie Sanders (@berniesanders) and Mayor Zohran Mamdani (@zohrankmamdani). For the Republican-leaning accounts, we followed President Donald Trump (@realdonaldtrump), Vice President J.D. Vance (@jdvance), Secretary of State Marco Rubio (@marcorubio), Secretary of De-

fense Pete Hegseth (@petehegseth), Governor Ron DeSantis (@rondesantis), Secretary of Health and Human Services Robert F. Kennedy Jr. (@seckennedy) and Director of National Intelligence Tulsi Gabbard (@tulsigabbard), among others (titles as of January 2025).

Regression Models

This subsection presents the model parameters, convergence diagnostics, and forest plots displaying the odds ratios (with 95% HDIs) for the models described in Section 3.

Model Parameters Unless otherwise specified, all models were fit using four Markov chains, each with 2000 posterior draws after tuning for 2000 steps at a target acceptance of 0.95. Results are reported at 95% highest-density intervals (HDIs). A fixed random seed of 42 was used to ensure reproducibility.

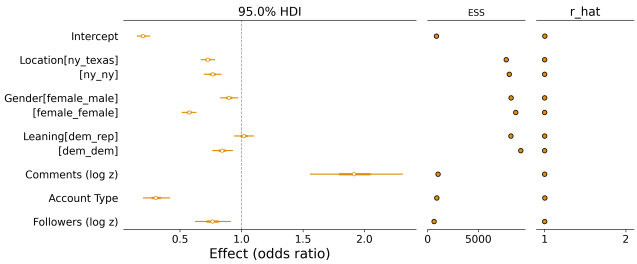

Figure 3: Forest plot of Bayesian logistic regression coefficients shown as odds ratios for Post-Level Regression. The dashed vertical line at 1 marks "no effect," and the right panels report diagnostics for each parameter.

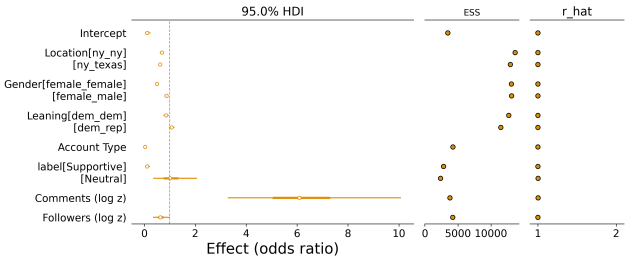

Figure 4: Forest plot of Bayesian logistic regression coefficients shown as odds ratios for Comment-Level Regression. The dashed vertical line at 1 marks "no effect," and the right panels report diagnostics for each parameter.

Annotation of comments using OpenAI API

The comments on each post from both news and non-news accounts were labeled 'Supportive/Against/Neutral' relative to the post. To do this, we first manually annotated 30 randomly selected comments. We then used the following prompt to obtain labels for all the comments. No modifications were made to the default API settings. Finally, validating LLM annotations against the manually annotated comments showed 0 mismatches.

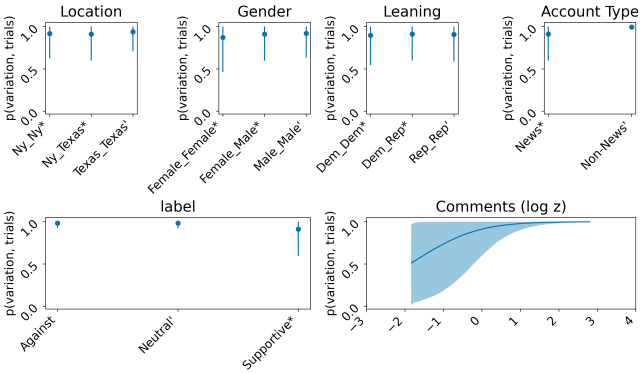

Figure 5: Posterior mean predictions for Comment-Level Regression across categorical and continuous covariates, holding other predictors constant. The categories with ' are the reference categories, and * indicates categories whose credible interval for the odds ratio excludes 1.

Prompt You are a researcher annotating Instagram posts. Given the caption of an Instagram post and a comment on that post, you are trying to determine the stance of the commenter with respect to the caption.

Label the comment as one of the following categories:

- Supportive: The comment expresses agreement, approval, or positive sentiment towards the content of the caption.
- Against: The comment expresses disagreement, disapproval, or negative sentiment towards the content of the caption.
- Neutral: The comment neither supports nor opposes the content of the caption; it may be factual or unrelated in tone.