

Moral Contagion or Emotional Echo? A Preregistered Replication of Brady et al. (#240762)

Author(s)

Niklas Bacher (University of Konstanz) - niklas.bacher@uni-konstanz.de

Pre-registered on: 2025/07/31 - 06:36 AM (PT)

1) Have any data been collected for this study already?

It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?

RQ1: Does moral-emotional language increase the upvote count of Reddit submissions in morally polarized subreddits?

H1: Reddit posts containing more moral-emotional language will receive more upvotes than posts with less moral-emotional content.

RQ2: Is the effect of moral-emotional language on submission diffusion stronger than the effect of purely emotional or purely moral language

H2: The effect of moral-emotional word usage on Reddit post score will be significantly stronger than the effect of purely moral or purely emotional language.

RQ3: Does the effect of moral-emotional language on submission diffusion differ across subreddits?

H3: The effect of moral-emotional language will vary across subreddits. Specifically, the effect will be weaker or even negative in subreddits with moderation norms encouraging balanced or deliberative discourse (e.g., r/NeutralPolitics) compared to subreddits with more polarized or emotionally charged content (e.g., r/IsraelPalestine, r/AbortionDebate).

3) Describe the key dependent variable(s) specifying how they will be measured.

Dependent Variable

- score: The Reddit post score, calculated as upvotes minus downvotes. This serves as a proxy for message diffusion, analogous to retweet counts in Brady et al. (2017). Note that Reddit's scoring system differs from Twitter's, as it integrates both positive and negative evaluations. However, as there are no negative scores, the score should work as well (and is applicable for a NB Model).

Independent Variables

The following three variables are mutually exclusive:

- moral_count: Number of moral words in the post text, based on the MFD dictionary. Z-standardized.

- emo_count: Number of emotional words in the post text, based on LIWC-style dictionary replicated from Burton et al. (2021). Z-standardized.

- moral_emo_count: Number of words that appear in both the moral and emotional dictionaries. Z-standardized.

Control Variable

- token_count: Total number of tokens (words) in the post text (selftext). Z-standardized. This is crucial due to the wide variance in Reddit post length, unlike Twitter.

4) How many and which conditions will participants be assigned to?

Not applicable

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Framework:

- All hypotheses will be tested using negative binomial regression models, with score (upvotes minus downvotes) as the outcome.

- The main predictors are counts of moral words, emotional words, moral-emotional words (intersection of the two lexica).

- The only control variable is the tokenized length of the post, due to data availability

- Posts with extreme length values (Token count > 10,000) are excluded as outliers to ensure convergence and avoid skewing the regression.

- Because Reddit submissions vary greatly in length compared to tweets (which used to be capped at the time Brady et.al did their analysis), I explicitly control for this structural difference by including token_count

Data consists of publicly available Reddit submissions from four subreddits. I selected the subreddits based on their thematic relevance to the original study and variation in discourse norms:

- r/Abortiondebate

- r/CovidVaccinated

- r/IsraelPalestine

- r/NeutralPolitics

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Inclusion:

- Only Reddit submissions (i.e.posts, not comments) from selected subreddits.

- Only posts in English (based on subreddit default language)
- Posts must contain textual content in the column selftext.

Exclusion:

- Posts with a token length (word count) exceeding 10000 are excluded as outliers due to their extreme leverage and to ensure model convergence in negative binomial regression.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

The data were obtained from a publicly available full Reddit data dump provided by the Pushshift project, specifically from this community post: https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/. The files are freely (but somehow tedious I have to admit) available via torrent and cover historical Reddit submissions for the top 40,000 subreddits from June 2005 to December 2024.

- r/AbortionDebate: 7,032 posts
- r/CovidVaccinated: 14,028 posts
- r/IsraelPalestine: 5,840 posts
- r/NeutralPolitics: 4,886 posts

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

Nothing else to pre-register.