Negative Expressions are Shared more on Twitter for Public Figures than for Ordinary Users

Jonas P. Schöne[1,2,3], David Garcia[4,6,7], Brian Parkinson [1], Amit Goldenberg [2, 3, 5]

[1] University of Oxford, Department of Experimental Psychology
[2] Harvard University, Harvard Business School
[3] Digital, Data and Design Institute at Harvard
[4] University of Konstanz, Department of Politics and Public Administration
[5] Harvard Department of Psychology
[6] Graz University of Technology, Department of Computer Science and Biomedical Engineering
[7] Complexity Science Hub Vienna

* **Corresponding Author:** Jonas P. Schöne[1]

Email: jonas.schone@psy.ox.ac.uk

Author Contributions: Jonas Schöne: Designed research; Performed research; Contributed new reagents/analytic tools; Analyzed data; Wrote the paper. David Garcia: Performed research; Contributed new reagents/analytic tools; Analyzed data. Brian Parkinson: Designed research; Wrote the paper. Amit Goldenberg: Designed research; Contributed new reagents/analytic tools; Analyzed data; Wrote the paper

Competing Interest Statement: On behalf of all authors, the corresponding author states that there is no conflict of interest. Authors declare no competing interests

Classification: Social and Political Sciences/Psychological and Cognitive Sciences

Keywords: Emotion Sharing, Social Media, Twitter, Emotion Contagion, Public Figures

Abstract

Social media users tend to produce content that contains more positive than negative emotional language. However, negative emotional language is more likely to be shared. To understand why, research has thus far focused on psychological processes associated with tweets' content. In the current study, we investigate if the content producer influences the extent to which their negative content is shared. More specifically, we focus on a group of users that are central to the diffusion of content on social media – public figures. We found that an increase in negativity was associated with a stronger increase in sharing for public figures compared to ordinary users. This effect was explained by two user characteristics, the number of followers and thus the strength of ties, and proportion of political tweets. The results shed light on whose negativity is most viral, allowing future research to develop interventions aimed to mitigate overexposure to negative content.

Significance Statement

Though most original social media content expresses positive emotions, content expressing negativity is shared more, inflating the platform's total negativity. In this study, we asked: whose negative content is more likely to be shared. Based on previous research, we suspected public figures' negativity is shared more due to the structure of their social network structure and their produced content. We found that negativity boosts the likelihood of public figures' content being shared, more than ordinary users. This is explained by the fact that negativity is more likely to be shared for weaker social ties and political content. Our findings offer insights into how negativity is shared and provide the basis for interventions reducing overexposure to negativity.

Main Text

Introduction

Most original content on social media is positive in affective tone (1-3). Yet, there is a growing realization that negative content is shared more than positive content (4-6). Users' increased tendency to share negative emotions inflates exposure to negativity on social media compared to its true proportion in content production. Overexposure to negativity is known to have adverse consequences at the individual level, leading to a reduction in well-being (7-9). At the collective level, exposure to negativity contributes to group polarization and intergroup conflicts (10, 11). Therefore, it is crucial to understand the roots of negativity sharing online and its driving mechanisms.

Previous research on negativity sharing has mainly focused on specific content-level features and psychological mechanisms that encourage the sharing of negative tweets (5, 12, 13). Here we hope to tackle the question in a complementary way by asking: *whose* negative content is more likely to be shared? More specifically, we hope to examine whether the association between negativity and sharing is stronger for public figures compared to ordinary users. We hypothesize that such association is stronger for public figures because they have distinctive characteristics that make their negative content more likely to be shared. Specifically, we show that two unique attributes seem to be responsible for the difference between public and ordinary users in the association between negativity and sharing. The first is the fact that public users have weaker ties, who are more likely to share negative emotions (14). The second is that public figures are more likely to write about politics, which is content that is typically more negative and more conducive to sharing (6).

To examine these hypotheses, we first replicated the previous finding showing original content on social media tends to be more positive. Second, we assessed whether the association between negativity and the number of retweets was stronger for public figures compared to ordinary users. We further examined, if this effect was driven by a certain type of public figures. We then compared two user characteristics, number of followers and proportion of political tweets, and assessed which of these user characteristics were associated with an increased likelihood that the negative content users generated was shared. Finally, we tested whether the differential effect of negativity on sharing for public figures and ordinary users was mediated by their distinctive user characteristics.

Negativity Sharing on Social Media

Shared content represents up to 75% of all content that people see on social media (15). It is therefore important to understand what context is more likely to be shared. Generally speaking, language that contains more emotional content is more likely to be shared (16), but one central question is whether positive or negative emotional language leads to more sharing. Although some studies have suggested that positive content, such as scientific articles (17), Olympic games posts (18), or news articles via email (19), is shared more than negative content, other studies have found that negative content tends to be shared more frequently than positive content in other contexts (4-6). The tendency to share negative content can be found in different cultures and platforms, including Facebook (5), Twitter (20), and Weibo (14).

Previous research has suggested several reasons why negativity might be shared more than positivity. The first reason is that heightened attention to negative content, also known as the negativity bias (21-23), may lead to more engagement and sharing (13, 24). The impact of the negativity bias seems to be moderated by tie strength (25), with negativity shared more between weaker ties, while positivity is shared more between close ties (14, 26). Given that negativity is more likely to be shared between weaker ties (25, 26), negativity should be more viral for users with a higher proportion of weak ties, such as public figures. A second reason why negativity is more likely to be shared is specific to political and intergroup discourse, which is frequent on social media (27-29). Users who write political tweets are often driven by intergroup hostility and reputation considerations, which might lead them to share more negative content (5, 6, 30). Therefore, users who are writing about politics more often may be more likely to have their negative content shared. It is important to note, however, that attention to political content and negativity sharing may also be driven by a bias in the literature towards political figures and news media. Recent research suggests that despite the fact that that Twitter users seem to be more engaged with politics than the average US population (31) the majority of users (60%) do not follow any political public figures on Twitter (32).

Public Figures and Ordinary Users

Verification status is one of the distinguishing features between users on major social media websites such as Twitter, Facebook, and Instagram. Verified users encompassed a wide range of public figures, including politicians, journalists, celebrities, and athletes. At the time of our data collection, the verification status on Twitter indicated whether a user was a public figure authenticated by the platform or not. The verified status of users changed on November 5th, 2022, after which every user was able to verify their account for $8 a month. Before the transition, verified users made up only a small proportion of all users. For example, Twitter has 229 million active users, of which only 420,300 (0.18%) were verified. Despite their relatively small number, verified users are central to the diffusion of content online (33, 34).

Public figures have distinctive characteristics on social media which may affect the extent to which their negativity would be associated with sharing. The first characteristic is their *high number of followers* (35), which often means that many of these ties are weak ties (36). Given that weaker ties are more likely to share negative content than positive content (14, 25), negative content generated by users who have many followers, such as public figures, is more likely to be shared compared to other types of content. The second characteristic of public figures on social media is that they produce a relatively *higher proportion of political content*. Public figures not only use social media to promote themselves but also to promote social and political causes (37-39). Additionally, many verified users on social media are political figures, journalists, or other users who specialize in politics, making them more likely to produce political content. Given that negativity is especially likely to be shared in political content (5, 6, 20), negative content from users who produce a higher proportion of political content is more likely to be shared.

Previous research has already demonstrated a positive link between negativity and increased content sharing when focusing on public figures such as political leaders or news outlets (20, 40, 41). However, these previous studies have not compared public figures to ordinary users and have not examined for whom the association between negative emotions and sharing is stronger. This question seems to be crucial if the ultimate goal is to find ways to reduce negativity sharing on social media. Furthermore, given that research on public figures has mostly focused on news media and political figures, it remains unclear whether the observed relationship between negativity and content diffusion can be generalized beyond this specific subset of public figures.

The Present Research

The primary goal of this study was to assess whether the association between negativity and content sharing is stronger when the content is produced by verified users compared to when it is produced by ordinary users. We further examined whether specific user characteristics – number of followers and proportion of political tweets – can account for this difference in the strength of association. To achieve these goals, we first tested whether the type of user (ordinary user, public figure) moderated the association between negative content and sharing. As there are various types of public figures who can obtain verification status, we further tested if certain types of public figures were more likely to be differentiated from ordinary users in the association between negative language and sharing. We then verified that public figures have relatively more followers and produce a higher proportion of political tweets, before investigating if these characteristics mediated the differential effect of expressed sentiment on sharing of content produced by public figures and ordinary users. The analyses were not pre-registered, but data and code are available at https://osf.io/xuraq/.

Results

Using the Twitter Application Programming Interface (API), we first compiled a list of users (n = 45,918), their account descriptions, and then extracted their tweets in January 2019 (see Materials and Methods section for more information on our data extraction procedure). We classified users into two the groups, public figures, and ordinary users, based on account verification status. At the time of our data collection, public figures had a blue checkmark indicating that their account was verified by Twitter, while ordinary users were not verified. Triangulating a few classification methods, we further classified public figures into categories, including Entertainment, Journalists, News Outlets, Organizations, Politics, Sport, and others. For a detailed description and breakdown of this classification, see the Materials and Methods section and the SI Section 5, Tables S10 – S12. We then selected an equal number of public figures and ordinary users (n = 6,678) who were matched by their activity level on the platform using propensity score matching. This method involved matching ordinary users to public figures based on their tweet count, as described in detail in the Materials and Methods section, resulting in a total sample of 427,502 tweets from public figures and 428,213 tweets from ordinary users (see Materials and Methods, and SI Section 1, Tables S1 – S3, and Figures S1 – S5 for analysis using the full sample). To assess user characteristics, we collected data on each user's number of followers as a measure of tie strength and analyzed their proportion of political tweets.

Following our process of user identification, we then turned to process the tweets produced by the users. A retweet occurs when one user shares another user's message with his or her own social network (43). For each tweet, we retrieved the number of retweets and evaluated the affective content of each tweet using the pre-evaluated sentiment analysis tool VADER (44). For each tweet, VADER generates a continuous sentiment score ranging from -1 (extremely negative) to +1 (extremely positive), along with an overall valence categorization (positive, neutral, negative). A tweet is classified as positive if the sentiment score exceeds 0.2, negative if it falls below -0.2, and neutral if the score falls between the two values. See the Methods section for a detailed explanation of the tweet evaluation process. We further compared the results of different sentiment analysis tools in SI Section 2, Table S4, and Figure S6. We identified political tweets using LDS topic modeling (42), which uses the co-occurrence of words or phrases to identify a pre-defined number of underlying themes. This method is capable of identifying political tweets because political content often contains similar words, such as the names of political figures or events. To identify the topic that represents political tweets, we manually inspected the most frequently occurring words in each topic and selected the one that contained political terms. More details on the assessment of these characteristics and their transformations in Materials and Methods and SI Section 3, Tables S5 – S9, and Figures S7 - S8 for details on the different configurations of the topic modeling.

Frequency of Positive, Negative, and Neutral Content for Public Figures and Ordinary Users. We tested if public figures and ordinary users produced more positive compared to negative, neutral affective content using the three VADER categories. For this analysis, we counted the number of a user's tweets in each of VADER's three affective categories. As we matched both user types by their total number of tweets, we were able to compare the absolute number of tweets in the given categories as the dependent variable. Using linear regression models, we predicted the total number of tweets per affective category based on user type (public figure, and ordinary user).

As expected, positive affective content was more frequent than negative for both user types (b = -12.79 [-13.78, -11.80], SE = 0.50, t (37358) = -25.40, p < .001, R^2 = .021, see Figure 1)[*]. The difference between positive and neutral content was only marginally significant (b = -0.90 [-1.88, 0.083], SE = 0.50, t (37358) = -1.79, p = .072, R^2 = .021). Looking at the comparison between public figures and ordinary users, we found that public figures produced a similar amount of positive content compared to ordinary users (b = 1.09 [-0.30, 2.48], SE = 0.71, t (37358) = 1.53, p = .12, R^2 = .021), a similar amount of neutral content (b = -1.32 [-3.29, 0.64], SE = 1.01, t (37358) = -1.31, p =.18, R^2 = .021), but most importantly less negative content compared to ordinary users (b = -2.15 [-4.12, -0.17], SE = 1.01, t (37358) = -2.13, p = .032, R^2 = .021). These results suggest that public figures tended to produce fewer negative tweets compared to ordinary users

* Marginal R^2 calculated based on recommendations from (45) using r package "MuMln" (46)

Associations between Sentiment Scores and Number of Retweets for Public Figures and Ordinary Users. To examine the association between negativity sharing and retweets for each user type, we conducted an interaction between user type (public figure and ordinary user) and the continuous sentiment score from VADER in predicting the log-modulus transformed the number of retweets (see Materials and Methods for more details on the transformation). According to previous research, an increase in both positive and negative sentiment should lead to more retweets, therefore we decided to fit a quadratic mixed model to predict the number of retweets using a quadratic function of the continuous sentiment score between -1 and 1. We also investigated potential non-quadratic relationships between sentiment and sharing but the quadratic model seemed to produce the most predictive model (see SI Section 7, Tables S14 – S15, and Figures S12 – S13). A quadratic function ($y = ax^2 + bx + c$) returns three coefficients describing the parabola. The coefficient a defines how wide the u-shaped graph is and if it opens upwards or downwards. If a is positive, then the parabola opens upwards, and if a has a higher absolute value, this means that the line slopes more steeply. Coefficient b represents whether and to what extent the local peak is a positive or a negative x value. If b is negative the local peak is a positive x value and if b is positive the local peak is a negative x value. Coefficient c is the intercept with the y-axis (at x = 0). A larger a coefficient in the present model would indicate a stronger association between sentiment and content sharing, while a larger b coefficient would suggest that positivity was shared more than negativity. In other words, the coefficient a informs us about the overall influence of emotional intensity or extremity on content sharing, while coefficient b informs us about whether positivity or negativity is more likely to be retweeted. To account for differences in the number of tweets produced by users in our model, we included a random intercept for each user.

Looking first at the interaction, results suggested that higher sentiment values (positive or negative) were more strongly positively associated with the number of retweets for content produced by public figures than for content produced by ordinary users (a = 0.18 [0.17, 0.19], SE = 0.0067, t (845141.16) = 27.17, $p < .001$, $R^2 = .085$, see Figure 2). This means that emotional content in general was more likely to be shared for public figures. The extent to which negativity led to more content sharing than positivity was also greater for public figures than for ordinary users (b = -0.12 [-0.12, -0.10], SE = 0.0036, t (845324.79) = -31.95, $p < .001$, $R^2 = .085$).

Having established the fact that the association between negativity and sharing was stronger for public figures, we then examined whether this effect was driven by certain types of public figures. To achieve this, we replicated the previous mixed model analysis, using the log-modulus transformed retweet count as our outcome variable and the interaction of two predictors: the quadratic function of the continuous sentiment score from VADER and the user type category. Unlike our previous analysis, where the user type variable was a simple binary variable (verified users vs. ordinary users), we expanded it to include multiple user types, including ordinary users and seven types of public figures. We also included a random intercept for individual users.

We found that negativity was shared more for all types of public figures compared to ordinary users (see SI Section 5, Table S12, and Figure S9 for full description), even in comparison to the subset of public figures for whom negativity was associated with the smallest increase in retweets, namely, organizations (*b* = -0.034 [-0.044, -0.023], *SE* = 0.0053, *t* (845324.79) = -6.37, $p < .001$, $R^2 = .053$). Negativity increased retweets most for political figures (*b* = -0.32 [-0.35, -0.30], *SE* = 0.012, *t* (845324.79) = -26.51, $p < .001$, $R^2 = .053$), followed by news outlets (*b* = -0.21 [-0.24, -0.16], *SE* = 0.020, *t* (845324.79) = -10.19, $p < .001$, $R^2 = .053$).

Differences in User Characteristics between Public Figures and Ordinary Users. We suspected that two variables can explain why negativity is more frequently shared for public figures: the difference in the number of followers and the proportion of political tweets. We first needed to establish that public figures indeed have a higher number of followers and a greater proportion of political tweets than ordinary users. We used a simple linear regression with a dummy-coded variable for verification status to predict the log-modulus transformed number of followers, and the proportion of their political tweets. The proportion of political tweets was defined as the number of political tweets identified by topic modeling divided by the total number of tweets.

Compared to ordinary users, public figures had more followers (*b* = 4.40 [4.34, 4.46], *SE* = 0.029, *t* (13352) = 151.3, $p < .001$, $R^2 = .63$, see Figure 3A), and produced approximately 2% more political content (*b* = 0.019 [0.014, 0.024], *SE* = 0.0025, *t* (13352) = 7.48, $p < .001$, $R^2 = .041$, see Figure 3B). While the proportion of political tweets only slightly differed between verified and ordinary users (around 2% increase), the difference in their number of followers was more substantial ($d = 0.01$ vs. $d = 0.16$). This suggests that follower count is likely a more salient factor in differentiating public figures from non-public figures than political tweet content.

Association between User Characteristics and Negativity Sharing. After confirming that public figures had more followers, and talked more about politics, we tested if these characteristics moderated the effect of sentiment on the number of retweets. To achieve this, we conducted two mixed models with quadratic terms. First, we looked at the interaction between the quadratic term of the continuous sentiment score and the number of followers in predicting the number of retweets. We also included a random intercept for users in both models. We found a stronger association between general sentiment and the number of retweets (*a* = 0.049 [0.083, 0.096], *SE* = 0.0033, *t* (846247.48) = 26.88, $p < .001$, $R^2 = .21$, see Figure 4), and more specifically between negativity and the number of retweets (*b* = -0.061 [-0.064, -0.057], *SE* = 0.0018, *t* (846530.08) = -33.29, $p < .001$, $R^2 = .21$) for tweets produced by users with more followers. Given the dramatically higher number of followers of verified users, we wanted to make sure that the effect of the number of followers is not limited to verified users. We, therefore, repeated this analysis using only the subsample of ordinary users, finding a similar effect (see SI Section 4). Additionally, we tested another model in which we matched a subset of ordinary users and public figures and based on their number of followers. We found that negativity was shared more for ordinary users that have as many followers as some public figures, although not to the same extent (see SI Section 4 for detailed discussion).

In the second model, we examined the interaction between the quadratic term of the continuous sentiment score and the proportion of political tweets. We again used a random intercept for users as in the previous models. We found that tweets that were produced by users with a higher proportion of political tweets showed a stronger association between general sentiment (positive or negative) and the number of retweets (a = 0.014 [0.0078, 0.021], SE = 0.0034, t (854258.36) = 4.23, p < .001, R^2 = .014, see Figure 5), and between negativity and number of retweets (b = -0.017 [-0.021, -0.013], SE = 0.0019, t (854302.00) = -9.17, p < .001, R^2 = .014).

Parallel Mediation Analysis. We hypothesized that negative content produced by public figures was shared more due to the differences in user characteristics that promote negativity sharing. To assess this prediction, we conducted a parallel mediation analysis assessing two potential mediators of the effects of user type (public figures vs. ordinary users) on sharing of their negative content (Hayes, 2017). To conduct the mediation, we needed an individual-level variable that reflected the degree to which negativity was shared for that user. We computed a new dependent variable that quantified the extent to which an increase in negativity was associated with more retweets for every individual user in our dataset, which allowed us to predict how much negativity is shared for certain users depending on their characteristics. This negativity-sharing dependent variable was calculated using a similar model to the previous ones with two additional changes. First, instead of a quadratic mixed model, we used a split regression to approximate the U-shaped relationship between sentiment and the number of retweets. A split regression contains a categorical variable that is inserted into a linear regression model as an interaction factor to allow for separate slopes for different categories. In our case, we split the continuous variable sentiment using a binary categorical variable into values < 0 (negative slope) and ≥ 0 (positive slope). This approach allowed us to derive a single coefficient specifically representing the association between an increase in negativity and the number of retweets. Second, we introduced random slopes representing the relationship between sentiment and the number of retweets for each user. Another benefit of using a split regression for the extraction of a per person coefficient was that a linear regression uses one coefficient to describe the relationship between sentiment and retweets (the beta coefficient), while a quadratic regression uses two (coefficients a and b, describing how much sentiment was associated with retweets as well as if negative emotions were shared more than positive emotions). We extracted the random slope describing the extent to which an increase in negative emotion was associated with the number of retweets as our dependent variable for the mediation analysis. The potential mediators were the two user characteristics identified above, namely the number of followers, and the proportion of political tweets. We used the PROCESS v4 macro for RStudio by Hayes (47) to conduct the parallel mediation analysis.

Starting with the a-paths in our parallel mediation model, user type was, as reported above, a significant positive predictor of both number of followers (a_1 = 4.40 [4.33, 4.45], *SE* = 0.029, *t* (13352) = 150.82, $p < .001$, $R^2 = .63$, see Figure 6), and the proportion of political tweets (a_2 = 0.019 [0.014, 0.024], *SE* = 0.0025, t (13352) = 7.47, $p < .001$, $R^2 = .0042$). Both the number of followers (b_1 = 0.019 [0.018 - 0.0021], *SE* = 0.0006, *t* (13352) = 31.01, $p < .001$, $R^2 = 0.13$), and the proportion of political tweets were also significant predictors of sharing of negative content (b_2 = 0.018 [0.043 - 0.0032], *SE* = 0.0072, *t* (13352) = 2.55, $p = .01$, $R^2 = 0.13$). While the total effect of user type on negativity sharing was significant (c = 0.081 [0.076, 0.085], *SE* = 0.0022, *t* (13352) = 37.08, $p < .001$, $R^2 = 0.13$), the direct effect controlling for the mediators rendered this effect nonsignificant (c' = -0.0047 [-0.011 - 0.0021], *SE* = 0.0035, *t* (13352) = -1.36, $p = .17$, $R^2 = 0.13$). A 95% bias-corrected confidence interval based on 10,000 bootstrap samples indicated that the sampled indirect effects of user type on sharing of negative tweets via the number of followers (a_1b_1 = 0.083, *SE* = 0.0039), while holding the other mediator constant, were consistently above zero (0.075 - 0.091). These findings indicate a significant positive indirect effect, suggesting that the number of followers mediates the relationship between user type and sharing of negative tweets. Similarly, the sampled indirect effects via the proportion of political tweets were also consistently above zero (a_2b_2 = 0.0004 [0.0001 - 0.0006], *SE* = 0.0010). These results from the parallel mediation analysis suggest that user type only had an indirect effect on negativity sharing. This effect was mediated both by the user's number of followers and proportion of political tweets.

Discussion

In this project, we compared the extent to which emotional content is shared for public figures and ordinary users. We found that despite the fact that public figures tended to produce less negative content than other users, the association between the increase in emotional intensity, and especially, negativity and the number of retweets the post received was stronger for public figures compared to ordinary users. This stronger association between negativity and sharing was consistent among all types of public figures, while we did not find that negativity was shared more than positivity for ordinary users. We identified two user characteristics – the number of followers and the proportion of political content – that mediated the effect of user type on the extent to which negativity was associated with an increase in retweets. When comparing these two mechanisms it seemed that the number of followers was a stronger mediator to the differences between the user types. This work supplements previous research on sharing of negativity which has mostly focused on psychological processes elicited by negative emotions on tweets (4, 6, 20).

Public figures seem to contribute substantially to people's exposure to negative content on social media. Whenever a tweet is shared it is duplicated and displayed to the sharer's followers. Considering the fact that public figures have a much larger number of followers and given their centrality in social media networks (33, 34), their shared content makes up a large share of the material presented on social media (see Figure 7). The resulting negatively biased sample of retweeted content then may lead other users to infer that the most credible and popular users on social media platforms use negative language, which in turn might negatively influence emotion-expression norms.

Disproportionate sharing of public figures' negative content could have adverse implications for both individuals and collectives on social media. Overrepresentation of negative information, such as negative news or online hate, cultivates a more negative evaluation of the world (48), potentially leading to a decrease in social trust (49), and a reduction in subjective well-being (50, 51). In addition, exposure to negative political content has negative collective consequences such as contributing to group polarization and intergroup conflicts (10, 11). This study's findings also provide an explanation for why the overrepresentation of negative emotional content produced by public figures has worsened over time (52). The increased sharing of negative affective content incentivizes public figures to generate more of it (53), thereby perpetuating the cycle of negativity on social media.

Limitations and Future Directions.

While this work provides new insights into how negativity is shared online, and despite our efforts to address alternative hypotheses, the current analysis has limitations that should be addressed in future work. The most important limitation is the observational nature of this study, which means that we could not manipulate user characteristics while controlling for others that differed between public figures and ordinary users such as average emotion expressed (see SI Section 6, Tables S13, and Figures S10 – S11 for the influence of average emotions expressed on negativity sharing). In future studies, researchers should manipulate user characteristics by using a curated news feed lookalike that allows for the manipulation of such user characteristics.

The second limitation pertains to our assessment of user characteristics. While it is plausible that average tie strength decreases as the number of followers a user has increased, we do not have a direct measure for tie strength, such as reciprocal connections or mutual interactions. In a similar vein, political content was classified using topic modeling which can be implemented in various ways. This raises the question of how accurately this classification can categorize political content. To alleviate some of these concerns, we tested different configurations of the topic modeling classifications, still finding similar results (see SI Section 3, Tables S5 – S9, and Figures S7 - S8). Future work should sample entire networks over time to measure how much users interact with each other to get a more fine-grained measurement of tie strength as well as a user's general tendency to create political content.

In addition to addressing the above-mentioned limitations, future research should seek to develop interventions designed to minimize general overexposure to negative content by targeting the above-mentioned user characteristics. We identified users whose negative content had the highest tendency to be shared. As negative content is in fact produced more rarely than positive content (2), an effective intervention should aim to prevent the followers of such users from disproportionally sharing negative content. One possible way of doing this is by educating users about the consequences of sharing negative content of users who have a high number of followers or by providing them with feedback about their tendency to share negative content produced by much-followed users (55). Given that the underlying mechanisms are likely driven by psychological tendencies, the increased consumption of negativity and the associated well-being risks may also manifest in other online contexts, such as browsing behavior (54) as well as in offline contexts such as or news consumption (40). As a result, solely intervening on social media may only address a fraction of the well-being risks associated with internet use.

We believe that our findings emphasize the crucial role that users with large followings such as public figures have in the dissemination of negative content. Furthermore, the findings shed light on the mechanisms that are involved in the process of sharing negative content and provide the basis for developing interventions aimed to combat the exposure of negative content online.

Materials and Methods

This research adhered to the best practice guidelines for Internet-mediated research set forth by the Central University Research Ethics Committee (CUREC) of the University of Oxford. According to these guidelines, the analysis of public data does not require further ethics approval.

Participants.

Based on previous research, we aimed to collect at least 350,000 tweets per user type to detect the effect of emotion expression on sharing (4, 6). We estimated that in one month, we could collect approximately 7,000 ordinary users (assuming that the median number of tweets is 50, 56). To collect a list of users from both user types, we used two separate approaches. For public figures, we first retrieved a full list of all public figures from the @verified Twitter account (N = 314,373) and their basic profile statistics. Next, we downloaded their tweets in the period of January 1st to January 31st, 2019. Our final sample included 2,246,068 tweets produced by 39,241 public figures. We then turned to ordinary users. Because there is no suitable method to sample random ordinary users directly, we extracted account names from randomly sampled tweets. We used the 1% Spritzer stream, a real-time stream of a random selection of 1% of all tweets, to collect random tweets between January 14th to February 13th, 2019, that were produced by ordinary users as indicated by the absence of verification status. We then obtained user names from the producer of these random tweets. After discarding duplicated users, we collected the profile statistics such as their number of followers of 6,681 users from this list and retrieved their tweets as well as their descriptive information of these tweets including the number of retweets produced in January 2019 (1,927,684 tweets) using the Twitter API. We removed all tweets that were not in English and non-original tweets, meaning that we removed retweets that did not contain their own added text, resulting in a final sample of 428,223 tweets produced by 6,678 ordinary users.

To achieve an equal sample size of users with similar Twitter activity, we used propensity score matching to match ordinary users to public figures based on their tweet count (57). This statistical technique helps address possible confounding factors in observational studies driven by inherent differences in samples. First, we calculated a propensity score for each user, indicating their likelihood of tweeting during the given month. Then, we employed the nearest neighbor method to match verified and ordinary users with similar propensity scores. This approach ensured that the two groups produced a similar number of tweets during the one-month period, reducing any potential distortions resulting from different tweeting behaviors. Each ordinary user was matched to one public figure with the closest tweet count. After matching, the sample included 6,678 users of each type, with 427,502 tweets from public figures and 428,213 tweets from ordinary users. We repeated the analysis from the result section using the full sample of public figures before matching, finding similar results (see Supplementary Information Section 1, Tables S1 – S3 and Figures S1 – S5).

Measures

Sentiment Analysis. We used the sentiment analysis tool VADER (44) to estimate the affective content of tweets. VADER was specifically developed for sentiment analysis in social media and is especially suited for short texts such as those posted on Twitter (58). For each tweet, VADER returns a categorization of the content's overall valence (positive, neutral, negative) as well as a continuous sentiment score ranging from -1 (extremely negative) to +1 (extremely positive). For the statistical analysis, we used the continuous sentiment score. We repeated the analysis using a different sentiment analysis tool (SentiStrength, 59), see SI Section 2, Table S4, and Figure S6 for more details).

Topic Modelling. To identify users who produced a higher proportion of political tweets, we first needed to distinguish political tweets from non-political tweets. We used Latent Dirichlet Allocation topic modeling to identify political content in tweets (LDA; 60, 61). LDA clusters text into a pre-defined number of topics representing distinct themes. This enabled us to assess the extent to which each sampled user produced political content. We conducted the topic modeling analysis in RStudio (version 4.0.2) using the "topicsmodel" package (62).

The specificity/generality of the topics that are identified depends on how many of them are preselected by the investigator. If the investigator decides to examine a small number of topics, topics modeling will use broad brush strokes to divide the content but ignore finer distinctions. By contrast, specifying a large number of topics can result in topics that are too specific for the particular research question. Choosing the number of pre-defined clusters is done to balance the specificity and interpretability of the created topics (63). Our goal in the topic number selection was to find one general political topic using the smallest number of topics possible to avoid having multiple, more specific political topics.

The meaning of a topic was assessed qualitatively by analyzing the words used most frequently in this topic (64). The frequency of a word in a topic is expressed in the β-score ("beta-scores"). After manual exploration of the semantic coherence of the topics, we found that using 5 topics created one topic that seemed to be almost exclusively about politics (as indicated by high β-scores for words such as "Trump," "president," "vote," "government" etc.; see SI Section 3, Tables S5 – S9, and Figures S7 - S8 for all efforts and details of the identified topics). After deciding on the number of topics, we derived γ-scores from the LDA analysis, which are percentage estimates of the likelihood that each tweet contained each of the specified topics. Based on this criterion, 25.03% of our sampled tweets were categorized as political tweets, which is similar to previous assessments of the quantity of political content on Twitter (65).

User-level Variables. Our user-level variables were user type, number of followers, and proportion of political tweets. A user was either categorized as the user type of public figure or ordinary user depending on whether the account was verified or not. Verification status and the number of followers were extracted from users' basic account information. However, the distribution of the number of followers was skewed and contained zero values. We, therefore, performed a log-modulus transformation (y=log(x+1)) on this variable before conducting our statistical analysis. The proportion of political tweets was calculated as the number of the user's tweets that were categorized as political by the topic modeling analysis (as described above) divided by their total number of tweets.

For verified accounts, we further classified them into several major categories of verified users, including political figures, journalists, news outlets, entertainment, sports, and organizations, and evaluated the tendency of their negative content to be shared by other users. To do so, we employed three classification approaches in conjunction to evaluate these categories. Our first approach was to analyze the most frequent words in users' profile descriptions, in order to identify potential categories and build manually curated dictionaries that describe the words used to classify users into their respective types (see SI Section 5, Tables S10 – S11 for all categories and corresponding word parts). Secondly, we matched users based on their identifiers with lists from previous research that had already classified them into specific public figure types. These lists include those created by Barberá (66) as well as Rathje, Van Bavel and van der Linden (5) for political figures, and the documentation by Bellovary et al. (20) for media outlets. Finally, we employed the tool "Demographer" (67), which utilizes machine learning and natural language processing techniques to infer whether an account belongs to an individual or organization from multilingual social media data.

Tweet-level Variable. We used the number of retweets as the main dependent variable. Because the distribution of the number of retweets was skewed and contained a high frequency of zeroes, we performed a log-modulus transformation before statistical analysis.

Acknowledgments

We are grateful for the helpful feedback on the manuscripts by Yan Yan (Stanford University) and Eric Shuman (New York University). This research was funded by the Promotionsförderung der Studienstiftung des deutschen Volkes and the Vienna Science and Technology Fund (Grant No. VRG16-005).

Preprints. This manuscript was published as a preprint at 10.31234/osf.io/wng5v.

Data Availability. The data and code used for this study are available on OSF at https://osf.io/xuraq/.

References

1. P. S. Dodds, *et al.*, Human language reveals a universal positivity bias. *Proc. Natl. Acad. Sci.* **112**, 2389–2394 (2015).
2. E. Ferrara, Z. Yang, Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Comput. Sci.* **1**, e26 (2015).
3. P. Baylis, *et al.*, Weather impacts expressed sentiment. *PloS One* **13**, e0195750 (2018).
4. W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci.* **114**, 7313–7318 (2017).
5. S. Rathje, J. J. Van Bavel, S. Van Der Linden, Out-group animosity drives engagement on social media. *Proc. Natl. Acad. Sci.* **118**, e2024292118 (2021).
6. J. P. Schöne, B. Parkinson, A. Goldenberg, Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affect. Sci.* **2**, 379–390 (2021).
7. E. Diener, *et al.*, New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Soc. Indic. Res.* **97**, 143–156 (2010).
8. P. E. Jose, B. T. Lim, F. B. Bryant, Does savoring increase happiness? A daily diary study. *J. Posit. Psychol.* **7**, 176–187 (2012).
9. L. Seidlitz, E. Diener, Memory for positive versus negative life events: theories for the differences between happy and unhappy persons. *J. Pers. Soc. Psychol.* **64**, 654 (1993).
10. V. O. de Mello, F. Cheung, M. Inzlicht, Twitter use in the everyday life: Exploring how Twitter use predicts well-being, polarization, and sense of belonging (2022).
11. W. J. Brady, *et al.*, Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nat. Hum. Behav.*, 1–11 (2023).

12. W. J. Brady, M. J. Crockett, J. J. Van Bavel, The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspect. Psychol. Sci.* **15**, 978–1010 (2020).

13. W. J. Brady, A. P. Gantman, J. J. Van Bavel, Attentional capture helps explain why moral and emotional content go viral. *J. Exp. Psychol. Gen.* **149**, 746 (2020).

14. R. Fan, K. E. Xu, J. Zhao, Higher contagion and weaker ties mean anger spreads faster than joy in social media. *ArXiv Prepr. ArXiv160803656* (2016).

15. D. Chang, G. Ghim, The structure and dynamics of the Korean Twitter network. *J. Commun. Res.* **48**, 59–86 (2011).

16. A. Goldenberg, J. J. Gross, Digital Emotion Contagion. *Trends Cogn. Sci.* **24**, 316–328 (2020).

17. K. L. Milkman, J. Berger, The science of sharing and the sharing of science. *Proc. Natl. Acad. Sci.* **111**, 13642–13649 (2014).

18. A. Gruzd, S. Doiron, P. Mai, Is Happiness Contagious Online? A Case of Twitter and the 2010 Winter Olympics in *2011 44th Hawaii International Conference on System Sciences*, (2011), pp. 1–9.

19. P. W. Kraft, Y. Krupnikov, K. Milita, J. B. Ryan, S. Soroka, Social media and the changing information environment: Sentiment differences in read versus recirculated news content. *Public Opin. Q.* **84**, 195–215 (2020).

20. A. K. Bellovary, N. A. Young, A. Goldenberg, Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly. *Affect. Sci.* **2**, 391–396 (2021).

21. S. Soroka, P. Fournier, L. Nir, Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proc. Natl. Acad. Sci.* **116**, 18888–18892 (2019).

22. S. N. Soroka, “Gatekeeping and the negativity bias” in *Oxford Research Encyclopedia of Politics*, (2016).

23. R. F. Baumeister, E. Bratslavsky, C. Finkenauer, K. D. Vohs, Bad is stronger than good. *Rev. Gen. Psychol.* **5**, 323–370 (2001).

24. H. Zhang, C. Qu, Emotional, especially negative microblogs are more popular on the web: evidence from an fMRI study. *Brain Imaging Behav.* **14**, 1328–1338 (2020).

25. L. K. Hansen, A. Arvidsson, F. Å. Nielsen, E. Colleoni, M. Etter, Good friends, bad news-affect and virality in twitter in *Future Information Technology: 6th International Conference, FutureTech 2011, Loutraki, Greece, June 28-30, 2011, Proceedings, Part II*, (Springer, 2011), pp. 34–43.

26. J. Berger, K. Milkman, Social transmission, emotion, and the virality of online content. *Whart. Res. Pap.* **106**, 1–52 (2010).

27. G. Ruhrmann, J. Woelke, M. Maier, N. Diehlmann, *Der wert von nachrichten im deutschen fernsehen: ein modell zur validierung von nachrichtenfaktoren* (Springer-Verlag, 2013).

28. S. Peterson, International news selection by the elite press: A case study. *Public Opin. Q.* **45**, 143–163 (1981).

29. D. M. Straughan, An experiment on the relation between news values and reader interest. *Gaz. Leiden Neth.* **43**, 93–107 (1989).

30. A. Goldenberg, *et al.*, Beyond emotional similarity: The role of situation-specific motives. *J. Exp. Psychol. Gen.* **149**, 138 (2020).

31. S. Wojcik, A. Hughes, How Twitter users compare to the general public. *Pew Res. Cent. Internet Sci. Tech* **156** (2019).

32. M. Wojcieszak, A. Casas, X. Yu, J. Nagler, J. A. Tucker, Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Sci. Adv.* **8**, eabn9418 (2022).

33. S. González-Bailón, M. De Domenico, Bots are less central than verified accounts during contentious political events. *Proc. Natl. Acad. Sci.* **118**, e2013443118 (2021).

34. H. Wang, K. Lei, K. Xu, Profiling the followers of the most influential and verified users on Sina Weibo in *2015 IEEE International Conference on Communications (ICC)*, (IEEE, 2015), pp. 1158–1163.

35. M. Cha, H. Haddadi, F. Benevenuto, K. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy. *Proc. Int. AAAI Conf. Web Soc. Media* **4**, 10–17 (2010).

36. S. Wies, A. Bleier, A. Edeling, Finding Goldilocks Influencers: How Follower Count Drives Social Media Engagement. *J. Mark.* **87**, 383–405 (2023).

37. M. Bossetta, The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 US election. *Journal. Mass Commun. Q.* **95**, 471–496 (2018).

38. K. Ross, T. Bürger, Face to face (book) Social media, political campaigning and the unbearable lightness of being there. *Polit. Sci.* **66**, 46–62 (2014).

39. N. E. Furlow, Find us on Facebook: How cause marketing has embraced social media. *J. Mark. Dev. Compet.* **5**, 61–64 (2011).

40. C. E. Robertson, *et al.*, Negativity drives online news consumption. *Nat. Hum. Behav.*, 1–11 (2023).

41. J. A. Fine, M. F. Hunt, Negativity and elite message diffusion on social media. *Polit. Behav.*, 1–19 (2021).

42. D. M. Blei, Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).

43. S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *J. Manag. Inf. Syst.* **29**, 217–248 (2013).

44. C. Hutto, E. Gilbert, 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*, (2014), pp. 216–225.

45. S. Nakagawa, P. C. Johnson, H. Schielzeth, The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* **14**, 20170213 (2017).

46. K. Barton, M. K. Barton, Package 'mumin.' *Version* **1**, 439 (2015).

47. A. F. Hayes, *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (Guilford publications, 2017).

48. G. Gerbner, L. Gross, M. Morgan, N. Signorielli, J. Shanahan, "Growing up with television: Cultivation processes" in *Media Effects*, (Routledge, 2002), pp. 53–78.

49. M. Näsi, P. Räsänen, J. Hawdon, E. Holkeri, A. Oksanen, Exposure to online hate material and social trust among Finnish youth. *Inf. Technol. People* **28**, 607–622 (2015).

50. T. Keipi, P. Räsänen, A. Oksanen, J. Hawdon, M. Näsi, Exposure to online hate material and subjective well-being: A comparative study of American and Finnish youth. *Online Inf. Rev.* **42**, 2–15 (2018).

51. M. Feinberg, B. Ford, S. Thai, A. Gatchpazian, B. Lassetter, The political is personal: Daily politics as a chronic stressor. *PsyArXiv Sept. 19 Doi* **10** (2020).

52. J. A. Frimer, *et al.*, Incivility is rising among American politicians on Twitter. *Soc. Psychol. Personal. Sci.* **14**, 259–269 (2023).

53. W. J. Brady, K. McLoughlin, T. N. Doan, M. J. Crockett, How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).

54. C. Kelly, T. Sharot, Knowledge-Seeking Reflects and Shapes Well-Being (2023).

55. E. Van Der Helm, M. P. Walker, Sleep and affective brain regulation. *Soc. Personal. Psychol. Compass* **6**, 773–791 (2012).

56. G. W. S. Ma, J. Schöne, B. Parkinson, Social sharing of emotion during the collective crisis of COVID-19 (2022).

57. E. A. Stuart, G. King, K. Imai, D. Ho, MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* (2011).

58. F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, F. Benevenuto, Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5**, 1–29 (2016).

59. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **61**, 2544–2558 (2010).

60. J. Berger, G. Packard, Are atypical things more popular? *Psychol. Sci.* **29**, 1178–1184 (2018).

61. D. M. Blei, J. Lafferty, *Topic models. Text mining: theory and applications* (Taylor and Francis London, 2009).

62. B. Grün, K. Hornik, M. B. Grün, Package 'topicmodels.' *Retrieved January* **1**, 2018 (2017).

63. C. Jacobi, W. Van Atteveldt, K. Welbers, Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digit. Journal.* **4**, 89–106 (2016).

64. J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, D. Blei, Reading tea leaves: How humans interpret topic models. *Adv. Neural Inf. Process. Syst.* **22** (2009).

65. Pew Research Centre, "The political content in users' tweets and the accounts they follow."

66. P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* **23**, 76–91 (2015).

67. Z. Wang, *et al.*, Demographic inference and representative population estimates from multilingual social media data in *The World Wide Web Conference*, (2019), pp. 2056–2067.

Figures Legends

Figure 1. *Number of affective tweets by categories for both user types.* The bar-graphs show that for both public figures and ordinary users negative affective content is the least frequent type of originally created content, replicating previous findings. Additionally, public figures seem to produce even less negative content compared to ordinary users.

Figure 2. *Number of retweets as a function of sentiment and user type*. The results suggest that stronger sentiment is associated with more retweets for both types of users. The local minimum for public figures is reached with a more positive emotional tweet, indicating that negativity is more strongly positively associated with number of retweets for public figures than for ordinary users. Public figures also received more retweets for neutral content than ordinary users.

Figure 3. *Differences between public figures and ordinary users.* Results suggest that public figures have more followers than ordinary users (Panel A) and produce a higher proportion of political tweets (Panel B).

Figure 4. *Number of retweets as a function of sentiment and number of followers*. To visualize the interaction between two continuous variables (sentiment and the number of followers), the panels show the mean number of followers in the middle panel as well as the *M* +/- SD (left and right panel respectively). The colored areas indicate the 95% Confidence intervals. The results show that the effect of sentiment on content sharing is greater when there is a higher number of followers. In other words, negativity sharing was stronger for the content of users with more followers.

Figure 5. *The number of retweets as a function of emotionality and proportion of political tweets*. To visualize the interaction of two continuous variables (sentiment and proportion of political tweets), the panels show the predicted association between sentiment and the number of retweets at the mean proportion of political tweets in the middle panel as well as the *M* +/- SD (left and right panel respectively). The colored areas indicate the 95% Confidence intervals. The results indicate that tweets produced by users with a higher proportion of political tweets show stronger associations between sentiment and content sharing and between negativity and content sharing.

Figure 6. *Parallel mediation analyzes the effect of user type on negativity sharing via two different user characteristics*. User type positively predicts the user characteristics: number of followers, and proportion of political tweets. The indirect effects of user type on negativity sharing via the number of followers and proportion of political tweets were both significant.

Figure 7. *Relative frequencies of emotional content in original tweets and retweets of public figures (A) compared to ordinary users (B).* Although negative content only made up 20.67% of the original tweets for public figures, it accounted for 30.92% of all retweets, signifying an increase of 10.25%. Conversely, for ordinary users, the proportion of negative content increased only slightly from 21.95% to 24.28% (2.33%). In contrast, when examining content that is less likely to be shared, such as neutral content for both user types, the proportion of such content decreases. Consequently, this content is underrepresented in retweets in comparison to its original frequency. These results suggest that the virality of negative content for public figures can lead to an inflation of their such content, compared to the original texts they produce.

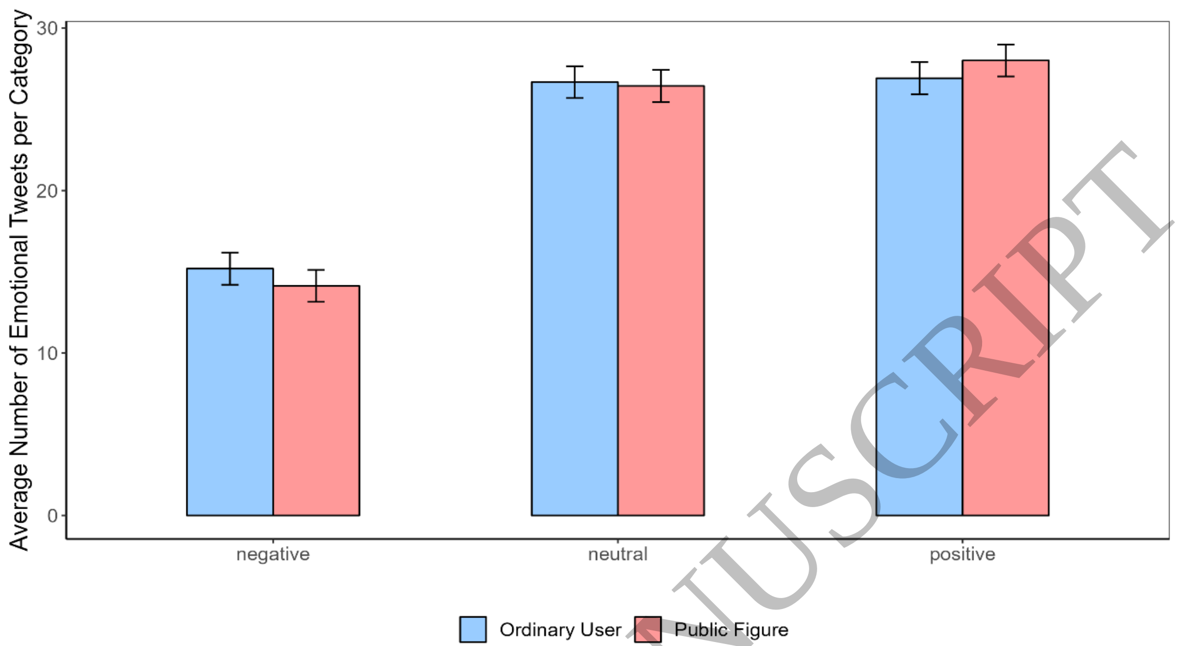


Figure 1
254x152 mm (x DPI)

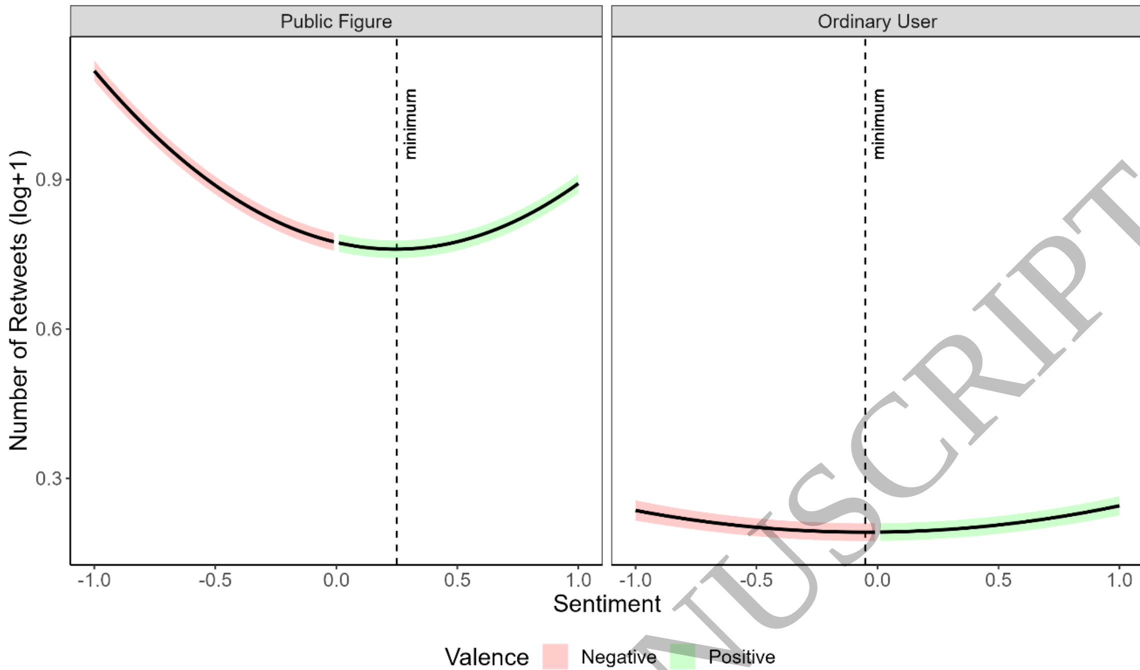


Figure 2
254x152 mm (x DPI)

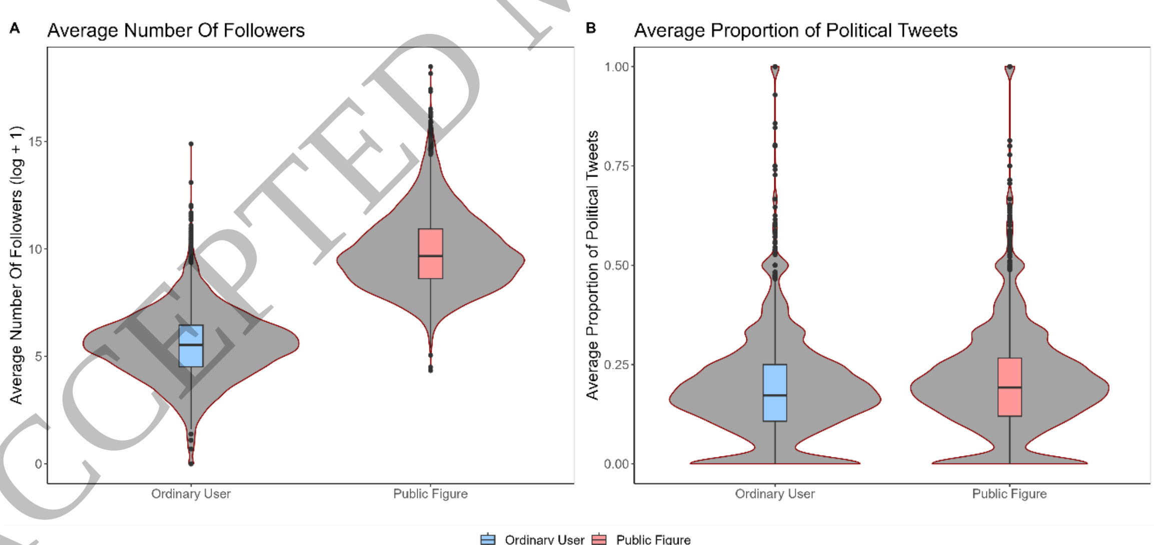


Figure 3
381x178 mm (x DPI)

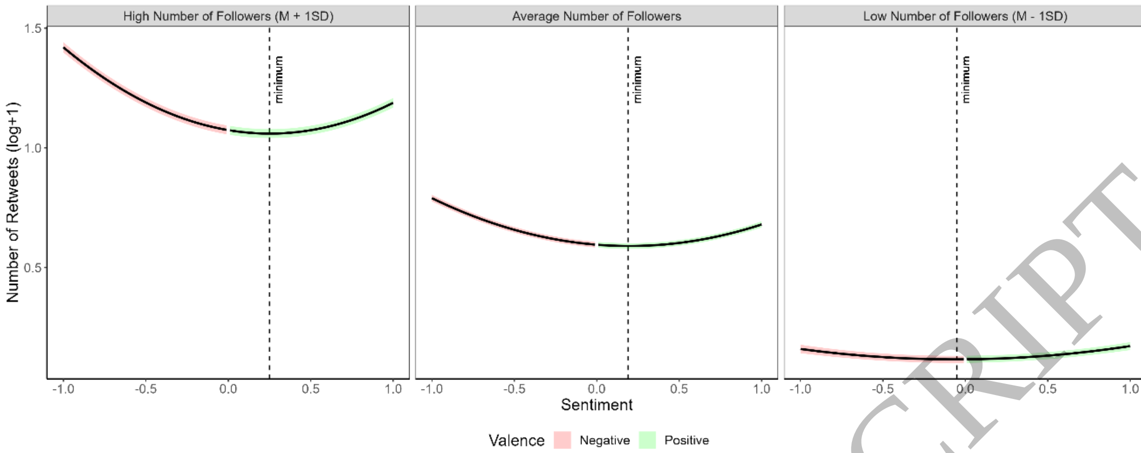


Figure 4
381x152 mm (x DPI)

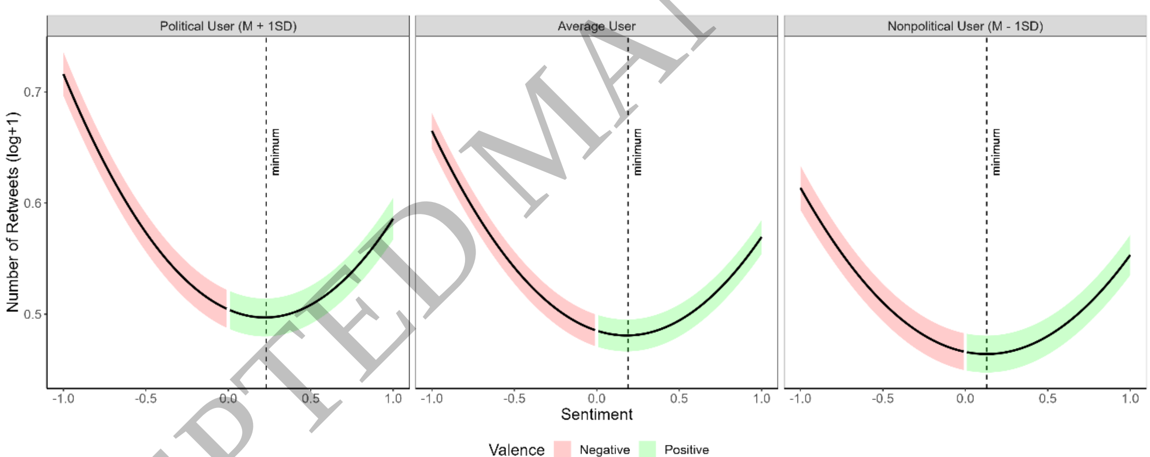
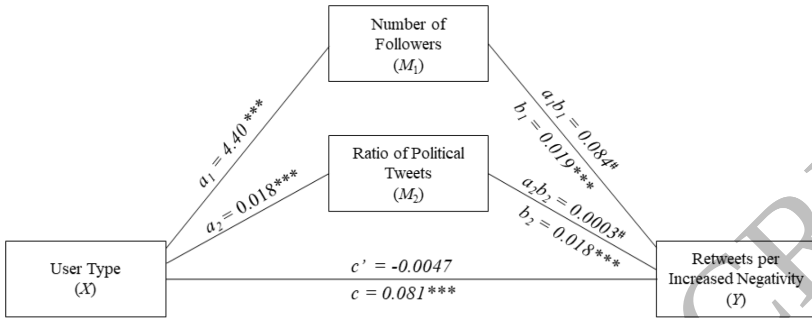


Figure 5
381x152 mm (x DPI)



1

2 *Figure 6*
3 *339x190 mm (x DPI)*