Article

new media & society
1–22

DOI: 10.1177/14614448251356240
journals.sagepub.com/home/nms
Sage

The consequences of "The Bird is Free": A computational analysis of aversive LGBTQIA+ tweets and engagement trends before and after Elon Musk dismantled the platform's moderation system

Gyo Hyun Koo
Howard University, USA

Josephine Lukito
The University of Texas at Austin, USA

Gina M Masullo
The University of Texas at Austin, USA

Christian Staal Bruun Overgaard
The University of Texas at Austin, USA

Bek Orr
State University of New York, Brockport, USA

Abstract

This research uses Twitter discourse about LGBTQIA+ people as a case study to investigate how changes in social media platform management and policies affect conversations about marginalized communities in digital spaces. We examine Twitter

Corresponding author:
Gyo Hyun Koo, Department of Communication, Culture and Media Studies, Howard University, 300 Bryant St NW, Washington, DC 20001, USA.
Email: hyun.koo@howard.edu

discourse before and after Elon Musk's acquisition and the immediate dismantling of content moderation efforts to identify changes in aversiveness in conversations about LGBTQIA+ people and users' engagement with such content. Time-series intervention analyses of 6 months of Twitter data ($n = 323,440$) show that tweets with toxicity, insults, and threats increased in the 3 months after Musk's takeover, although the differences were small. The volatility of aversive content also grew over time. Post-Musk, tweets with severe toxicity and insults received more engagement than before. This research provides insight into Twitter as one representation of the shared reality of today's cultural moment, with implications for minoritized groups and the free flow of information on social media platforms.

Keywords
Aversive online content, computational analysis, content moderation, Elon Musk, LGBTQIA+, online discourse, platform governance, social media engagement, time series analysis, Twitter

This research uses Twitter discourse about LGBTQIA+ people as a case study to investigate how changes in social media platform management and policies affect conversations about marginalized communities in digital spaces. We examine Twitter discourse before and after Elon Musk's acquisition and immediate dismantling of content moderation efforts to identify changes in aversiveness in conversations about LGBTQIA+ people and users' engagement with such content. Time-series intervention analyses of 6 months of Twitter data ($n = 323,440$) show that tweets with toxicity, insults, and threats increased in the 3 months after Musk's takeover, although the differences were small. The volatility of aversive content also grew over time. Post-Musk, tweets with severe toxicity and insults received more engagement than before. This research provides insight into Twitter as one representation of the shared reality of today's cultural moment, with implications for minoritized groups and the free flow of information on social media platforms.

Billionaire Elon Musk tweeted "the bird is free"[1] when he took over Twitter and immediately dismantled much of its content moderation. Musk, who has since rebranded Twitter as X,[2] laid off thousands, including engineers who criticized him and content moderators tasked with policing the platform's content (Perrigo, 2023). He also revoked suspensions of dozens of accounts that had violated the platform's terms of service, including then-former US President Donald J. Trump's account (Wile, 2022). Trump was banned from the platform, "due to the risk of further incitement of violence" (Fung, 2021), following an effort by some of his supporters to take over the US Capitol on 6 January 2021. Journalists, pundits, and scholars expressed concern that Musk's dismantling of content moderation and other safety mechanisms aimed at preventing hate speech would allow *aversive* content, defined as violating conversational norms of minimizing conflict and promoting accord (Papacharissi, 2004), to proliferate on the platform (e.g. Counts and Nakano, 2023; Wang et al., 2024; Yurcaba and Ingram, 2023). What little research there is addressing how Twitter changed after Musk took over has found that, in

general, hate speech increased on Twitter after Musk took over (e.g. Benton et al., 2022; Hickey et al., 2023). Little empirical work has considered whether Twitter became less hospitable for minoritized groups, which are most at risk for online hate (Kostić and McGonagle, 2019; Sobieraj, 2020).

In this study, we examined the effect of Musk's Twitter takeover and immediate dismantling of much content moderation on Twitter content through the lens of one minoritized group that Musk had expressed particular animosity toward: LGBTQIA+ people. Musk reinstated Twitter accounts that had been banned for anti-LGBTQIA+ harassment, mocking pronoun usage, and misgendering or deadnaming queer and trans people (Yurcaba and Ingram, 2023). Musk's own daughter, who is transgender, has repeatedly described his anti-queer rhetoric (Ingram, 2024). In addition, Musk's takeover of Twitter took place during a period when more than 500 bills[3] had been proposed across the United States to suppress LGBTQIA+ rights through acts, such as limiting gender-affirming care for minors, dictating what restrooms people can use, and prohibiting drag performance. These bills may foster an anti-LGBTQIA+ chilling effect by making a cage of the gender binary and punishing those who do not conform.

Through a computational analysis of tweets ($n = 323{,}440$) about LGBTQIA+ discourse, we compared the 3 months before Musk took over the platform on 27 October 2022 (from 27 July 2022 to 26 October 2022) to tweets in the 3 months after (from 27 October 2022 to 27 January 2023). We found that some types of aversive content, such as insults, significantly increased in Twitter discussions about LGBTQIA+ people, although changes were small. Other types of content, such as attacks on identity, did not change, and profanity decreased. We also found that users were more likely to engage with some types of aversive content after Musk took over Twitter, which has the potential to amplify aversive content. In addition, we found that the volatility of all aversive content increased over time, suggesting that aversive content increases and decreases more after Musk took over Twitter, compared to before. We use the case of discourse about LGBTQIA+ people as an exemplar of the larger societal problem of online discord in the wake of Musk's dismantling of content moderation on the Twitter platform. Notably, our data collection took place before Twitter discontinued its API open access for researchers (XDevelopers, 2023), adding challenges to studying online discourse. This marks our study as a valuable resource for understanding how the role of media ownership not only extends beyond the establishment of content moderation policies but also profoundly influences the wider public discourse and intervenes in the everyday lives of historically underserved communities.

Theorizing Twitter as ritual communication

Drawing from Carey's (1989) ritual communication theory, we theorize Twitter discourse as a form of *ritual* communication in that it can represent or even create a cultural moment, not just transmit information. The ritual communication viewpoint proposes that communication is about "construction and maintenance of an ordered, meaningful cultural world" (pp. 18–19). This approach focuses on the larger meaning communicated through a message. In contrast, the *transmission view* focuses on communication as the sharing of information, much like transportation. The transmission view is more

functional, focused on getting information *to* someplace (whether virtual or physical), whereas the ritual view interprets communication as a "representation of shared beliefs" that constructs and maintains a "cultural world" (Carey, 1989: 18–19). Ritual communication "is directed not toward the extension of messages in space but toward the maintenance of society in time" (Carey, 1989: 18). In other words, the ritual viewpoint sees communication as evidence of a shared cultural moment (Carey, 1989: 18). In our study, we see tweets operating under the ritual view of communication because they offer insight through their collective representation of shared cultural beliefs (Carey, 1989). In other words, our corpus of tweets presents a collective representation of public content about LGBTQIA+ people, offering insights about the larger discourses emanating through society. While Twitter is certainly not used by everyone, it represents a form of virtual public sphere where people engage in a "cultural conversation" (Brock, 2012: 59) and communicate their beliefs in a limited form of public opinion (McGregor, 2019). Indeed, journalists (McGregor, 2019), political groups, and social movements (Karpf, 2018) have used Twitter discussions to understand public sentiment and to gather information as they make decisions (Brock, 2012, 2018). Twitter is a space for the potential enactment of politically expressive talk that can influence people's political engagement (Min, 2007). Using Carey's (1989) conception, we argue that Twitter conversations can operate as a form of ritual communication that may represent a piece of the "shared beliefs" and "cultural world" (Carey, 1989: 18) in the United States in regard to LGBTQIA+ people. Thus, our interest is in what the collection of tweets as a corpus represents as a form of ritualized communication and what this explains about our culture and media. As a result, our focus is on any tweets that discuss LGBTQIA+ people or issues as evidenced by keywords in the content, although our data does not parse whether these content producers are from within or outside LGBTQIA+ communities. However, collectively this corpus offers insight into the public zeitgeist regarding LGBTQIA+ people.

Aversive online content

Because of Twitter's potential role as a form of public opinion, the platform has become a focal digital space to understand aversive content (e.g. Jikeli and Soemer, 2023; Klein, 2019; Oz et al., 2018; Poole et al., 2021; Rogers, 2020; Sobieraj, 2020). Aversive content is a sustained problem online, with as much as 20% of content fitting definitions of various types of aversive content, including incivility or toxicity (Chen, 2017; Coe et al., 2014), or even more for controversial topics (e.g. Harlow, 2015). Much evidence suggests that aversive content can have troubling implications, including causing emotional upset and anger (Chen and Lu, 2017; Gervais, 2015; Riedl et al., 2020). Based on this research on aversive content, we considered that content on Twitter might become more aversive toward LGBTQIA+ people after Musk acquired the platform, fired content moderators, and dismantled content moderation efforts. Our rationale comes from research showing that aversive content is the very type of content that content moderators and others typically remove (Vargo et al., 2024). Thus, it stands to reason that decreasing content moderation efforts and firing content moderators would make the

platform less hospitable to LGBTQIA+ people because they are one of the minoritized groups who are most at risk online (United Nations, 2021).

In our study, we examined six types of aversive content identified by Google's Perspective API—the algorithm developed by Jigsaw and Google's Counter Abuse Technology. It assesses the level of six types of aversive content in text and has been widely used in communication research (e.g. Frimer et al., 2022; Vargo and Hopp, 2020). The attributes are identity attacks, hateful content that targets someone's identity (e.g. gender or race); insults, negative inflammatory content directed toward a person or group; profanity, curse words, or obscene language; threats, content that expresses an intention to inflict injury or violence against a person or group; toxicity, rude, disrespectful, or unreasonable content; severe toxicity, content that is very hateful, aggressive, or disrespectful.[4] As with any general classification tool, Perspective API is imperfect; however, it can be useful when studying online comments (e.g. Salminen et al., 2020) and has the advantage of distinguishing different forms of aversive content.

We focused exclusively on Twitter, rather than comparing it to other platforms, because Musk's takeover was the first time a major social media platform had experienced such upheaval, so it offers a singular example of how aversive speech that targets marginalized groups may grow when content moderation is curtailed. In addition, Musk said part of the reason he purchased the platform was to reduce moderation and restore what he considered "free speech" to the platform (Dang, 2022, n.p.; see also Perrigo, 2023), so Twitter offers an opportunity to assess how content changes when content moderation is reduced.

Almost immediately after Musk's takeover of Twitter, there was a flurry of news reports about how Twitter would become more aversive because of Musk's policies (e.g. Counts and Nakano, 2023; Yurcaba and Ingram, 2023). Conservatives initially welcomed Musk's takeover because they hoped it would reverse what they perceived as a trend of silencing conservative voices on the platform, but Democrats decried it, claiming the changes Musk made could make the platform dangerous for democracy (Popli, 2022; Rohlinger et al., 2023) by removing limitations on speech that could lead to chaos. We focus specifically on Twitter, rather than comparing it with other platforms, because we are interested in understanding specifically how Musk's takeover changed content on that platform.[5]

Both critical and queer theories support the idea that members of minoritized groups would be expected to face more challenges online because their existence challenges dominant power structures, and racism, sexism, and heterosexism is embedded in US society (e.g. Bronski, 2011; Butler, 1999; Delgado and Stefanic, 2012; Foucault, 1985; Hall, 2000; Tawa and Bunts, 2022). Empirical data support that minoritized groups are more targeted in society. Queer people are almost nine times more likely to experience a hate crime than straight people (Flores et al., 2022). This literature underscores our focus on a specific minoritized group online as a case study to examine how content on Twitter changed in the wake of Musk's takeover.

Because no research has examined aversive content in LGBTQIA+ discourse after Musk's Twitter takeover, we first ask:

RQ1: Following Elon Musk's acquisition of Twitter, how did the aversiveness of tweets about LGBTQIA+ people change in terms of (a) toxicity, (b) severe toxicity, (c) identity attacks, (d) insults, (e) profanity, and (f) threats?

We also considered whether people on Twitter would engage more with aversive LGBTQIA+ content after Musk's takeover. We were specifically interested in engagement overall as a form of attention to aversive content, not in the specific attributes of engagement. Our rationale was that aversive content, in general, may elicit engagement (Muddiman and Stroud, 2017). Literature in platform studies, mass communication, and political science suggest that Musk's takeover may have motivated contentious actors (Barrie, 2023; Wang et al., 2024).

We would, therefore, expect aversive content about LGBTQIA+ people to follow this same pattern. However, it is unknown whether people would be more likely to engage with this content before the takeover or after. On one hand, if Twitter has become less hospitable to minoritized groups, including LGBTQIA+ people, after Musk's takeover, then it would be logical that people on the platform might feel emboldened to engage with aversive content because it is more plentiful, and the atmosphere is more free-flowing. But it is also plausible that there would be no change or that engagement with aversive content could decrease after Musk's takeover because the takeover might induce a chilling effect on engagement or change content in unanticipated ways.

RQ2: Is there a difference in levels of engagement with aversive tweets about LGBTQIA+ people after Elon Musk took over Twitter, compared to before?

Methods

Data collection

All the tweets in our sample contain LGBTQIA+-related content. We used the {academictwitteR} package in R (Barrie and Ho, 2021), to access the Twitter API to gather all English-language tweets that contained at least one of 151 query terms representing LGBTQIA+ identities, such as "lesbian," "gay," and "bisexual," and derogatory terms. These terms were compiled based on literature (e.g. Borgogna et al., 2019; ElSherief et al., 2018; Ribé et al., 2021), online encyclopedias (e.g. Wikipedia), and nonprofit organizations (e.g. Glaad, n.d.). Two queer authors and an additional queer scholar reviewed the list. The complete list of keywords and the frequency of each term are available in Online Supplemental Materials S1.[6] The query was not case-sensitive, and we excluded promoted tweets and retweets. Our time frame extended from 27 July 2022 to 27 January 2023, covering 3 months preceding (27 July 2022 to 26 October 2022) and 3 months following (27 October 2022 to 27 January 2023) Musk's acquisition of Twitter (N=32,713,090). For our analysis, we randomly selected about 1% of the sample (n=323,440).[7]

Table 1. Descriptive statistics of aversive content and engagement metrics.

	Before ($n = 150,521$)		After ($n = 172,919$)		Total ($N = 323,440$)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Aversive content						
Toxicity	0.3815441	0.21	0.3855321	0.21	0.3836762	0.21
Severe toxicity	0.09423872	0.15	0.09443035	0.15	0.09434117	0.15
Identity attacks	0.3310561	0.19	0.3335822	0.19	0.3324066	0.19
Profanity	0.2620881	0.21	0.2573377	0.21	0.2595484	0.21
Threats	0.03833399	0.08	0.0419311	0.08	0.0402571	0.08
Insults	0.245941	0.19	0.2548019	0.20	0.2506782	0.20
Engagement metrics						
Retweets	2.95	124.19	2.72	78.90	2.82	102.50
Replies	0.98	22.49	1.19	38.59	1.09	32.12
Likes	21.17	838.76	20.06	549.22	20.58	699.05
Quote retweets	0.38	23.41	0.24	8.15	0.30	17.05
Total Engagement	25.47	983.95	24.21	640.29	24.79	818.38

Unlike other descriptive statistics, we did not round the means of six aversive content scores to the second decimal place. This is because they range from 0 to 1, and rounding could obscure the differences between values, particularly in the "before" and "after" comparisons. The date 27 October 2022 is included in the "after" category. Using a threshold of 0.70, before Musk's acquisition of Twitter, 8.83% of tweets were classified as toxic: severe toxicity (0.27%), identity attack (1.36%), insult (1.72%), and profanity (6.59%). After the acquisition, the overall toxicity rate remained unchanged at 8.83%, with severe toxicity decreasing to 0.14%, identity attack increasing to 1.42%, insult rising to 2.05%, and profanity declining to 6.19%. Across the full stratified sample, 8.83% of tweets were classified as toxic, with rates of severe toxicity (0.20%), identity attack (1.39%), insult (1.90%), and profanity (6.37%). This threshold was chosen based on the literature, although it varies across different studies. For example, Pozzobon et al. (2023) used $\geqslant$0.5 and Weber et al. applied 0.38 for hate speech detection, 0.5 for accuracy, and 0.8 for precision. Nogara et al. (2023) employed thresholds ranging from 0.5 to 0.7.

Aversive content. To evaluate each tweet, we used the R package {peRspective} (Votta, 2021) to apply Perspective API's toxicity classifier, which generates a score (from 0 to 1) for the six dimensions of aversive content conceptually defined in the Literature Review: toxicity, severe toxicity, identity attacks, insults, threats, and profanity. Neither the literature nor the API itself specifies a consistent cutoff for when content is considered aversive; indeed, research has used cutoffs as low as 0.38 for some types of aversive speech (e.g. Weber et al., 2025) or as stringent as 0.70 (e.g. Nogara et al., 2023; Pozzobon et al., 2023). We employ the most conservative cutoff from the literature, and define aversive content as having a score of .70 or greater. We include several examples of aversive tweets in the online supplementary materials (S4), noting that some content may be stigmatizing or distressing. Table 1 presents descriptive statistics.

Engagement. We calculated the average of "retweets," "replies," "likes," and "quote retweets" ($M = 25.47$, $SD = 983.95$; see Table 1).

Data analyses

To answer RQ1, we conducted a time series intervention analysis using the {CausalImpact} package (Brodersen et al., 2015; for applications of this package, see Olteanu et al., 2018; Cervantes and Rambaud, 2020).[8] In time series, intervention analyses are used to understand whether and how a time series' data-generating process changes after an intervention. While there are several ways that a time series can change, our question is most interested in permanent and sustained increases of aversive content, as measured using the Perspective API. CausalImpact conducts this analysis by estimating a Bayesian structural model from the time series, accounting for both linear trends and seasonal cycles. Using this temporal structural model, it then estimates the time series if the data from the date of the intervention was not included, which allows researchers the ability to compare the actual time series (with the intervention) from a counterfactual (what would have happened if the intervention did not occur). When interpreting this analysis, the result with the intervention is the actual data, whereas the result without the intervention is the counterfactual estimated model. We conducted one intervention analysis for each dimension of the Perspective API.

The univariate descriptive analyses and aforementioned intervention analysis also indicated that it would be helpful to conduct a post hoc time series analysis to better understand how the variance of the time series changed before and after Musk's acquisition of Twitter. To do this, we use a GARCH model (Lundbergh and Teräsvirta, 2002). GARCH stands for "Generalized Autoregressive Conditional Heteroskedasticity," and it is a model for estimating whether volatility increases or decreases over a time period (Francq and Zakoian, 2019). Volatility is a time series property that indicates significant fluctuations in a time series; in the case of our data, this would result in high amounts of aversive content on some days, followed almost immediately by low amounts of aversive content on the days after. A common time series approach to modeling volatility is a GARCH model, which is widely used to measure the volatility of stock markets (e.g. Franses and Van Dijk, 1996; Salisu and Gupta, 2021).

To answer RQ2, we conducted a negative binomial regression analysis to predict engagement. Negative binomial regression was used because the outcome variable (engagement) is over-dispersed (Guo and Sun, 2020; Heiss et al., 2019). We excluded rows with zero engagement as suggested by Hilbe (2011). This reduced the sample to 215,365. We then used the {MASS} package (Ripley et al., 2023) to run a stepwise regression using the stepAIC function with "both" directions, combining both forward selection and backward elimination approaches. This removes predictors that have minimal impact on the Akaike information criterion (AIC). Based on this, we identified the best model with factors predicting "engagement" from six possible variables while avoiding overfitting. As a result, we reduced the predictors to three (toxicity, severe toxicity, and insults).

For the negative binomial regression, we included engagement as the outcome variable and entered the three predictors along with the time (date) variable. We coded the date of tweets posted as a categorical variable, with categories "before" (coded 0) and "after" (1) indicating posts published before and after Musk took over Twitter.[9] In addition, we added interaction terms to the regression model to assess any relationships between the time and each of the three dimensions, specifically "time x toxicity," "time

Table 2. Intervention analysis results.

	Actual	Predicted (*SD*)	Relative effect (*SD*)	CI	Posterior tail-area probability *p*
Toxicity	35.88	35.49 (0.12)	1.12% (0.35%)	0.39%, 1.8%	0.002
Severe Toxicity	8.85	8.78 (0.08)	0.89% (0.98%)	-1%, 2.6%	0.186
Identity Attacks	31.02	30.79 (0.12)	0.76% (0.41%)	-0.06%, 1.6%	0.035
Insults	23.68	22.88 (0.11)	3.51% (0.51%)	2.5%, 4.5%	0.001
Profanity	24.09	24.39 (0.12)	-1.2% (0.49%)	-2.2%, -0.25%	0.007
Threats	3.83	3.57 (0.03)	7.34% (1%)	5.4%, 9.3%	0.001

For the actual, predicted, and relative effect, the mean is reported with the standard deviation in parentheses. CI: confidence interval.

x severe toxicity," and "time x insults," in predicting engagement. Upon checking the variance inflation factor (VIF) of each predictor, there were no multicollinearity issues.[10]

Results

RQ1 asks how the aversiveness of Twitter content about LGBTQIA+ people shifted following Musk's purchase of the platform. We use a time series intervention analysis to compare the time prior to Musk's purchase with the period following Musk's purchase. An intervention analysis explores the effect of a change and compares the effects of the intervention to a counterfactual. We selected the day Musk's acquisition of Twitter as the intervention, as the platform dramatically changed before and after he took over. The counterfactual that is estimated, therefore, is what would have happened to the levels of aversive content in LGBTQIA+ discourse if Musk had not purchased Twitter. These results are summarized in Table 2.

In the case of toxicity (RQ1a), we find a small but statistically significant difference between the cumulative predicted levels of toxicity in the counterfactual ($M=35.49$, $SD=0.12$) and the actual effects of the intervention ($M=35.88$; as this is the actual effect of the intervention, rather than an estimate, there is no standard deviation to report), suggesting that toxicity increased following the intervention. While small (the relative effect of this is 1.1%), this result is statistically significant (Posterior tail-area probability $p=.002$)[11] and the relative effect of the intervention is within a 95% confidence interval [0.39%, 1.8%].

In the case of severe toxicity (RQ1b), we do not find a significant difference (posterior tail-area probability $p=.186$) between cumulative predicted levels of severe toxicity in the counterfactual ($M=8.78$, $SD=0.08$) and the effects of the intervention ($M=8.85$). This is confirmed by the crossing of the 95% confidence interval [−1%, 2.6%].

In the case of identity attacks (RQ1c), we likewise do not find a significant difference between cumulative predicted levels of identity attacks in the counterfactual ($M=30.79$, $SD=0.12$) and the effects of the intervention ($M=31.02$). While the posterior tail-area

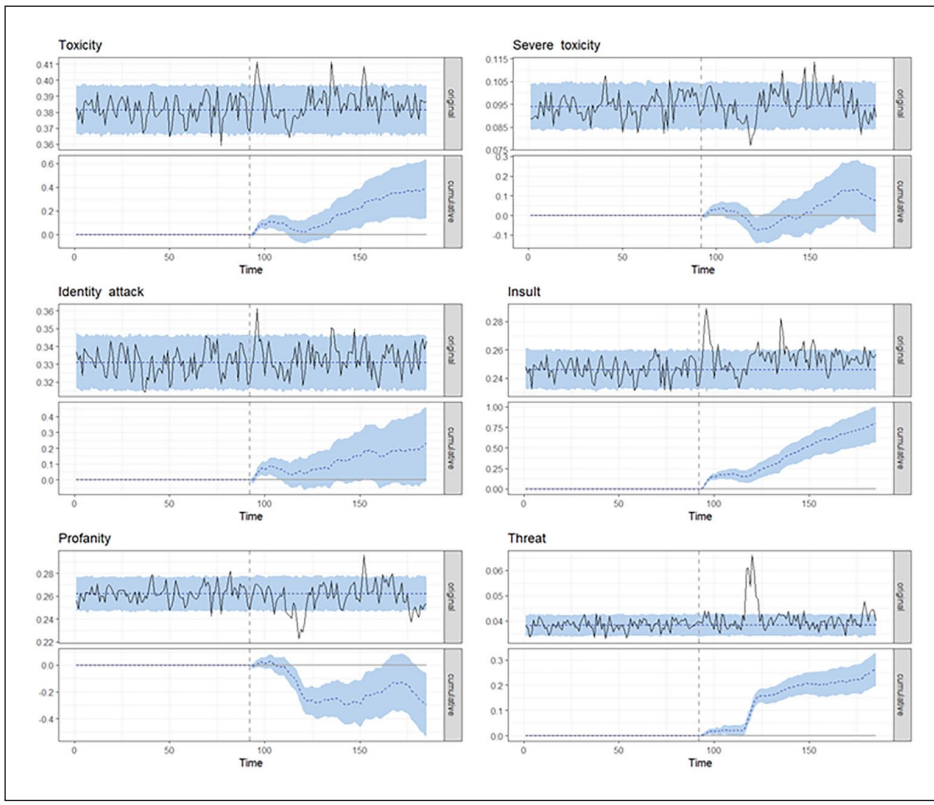

Figure 1. Effect of Elon Musk's Twitter acquisition on LGBTQIA+ aversive language.
Shades around the lines indicate confidence intervals. We also present a graph in the Online Supplemental Materials (S3) illustrating the number of tweets per day. Presented here are the cumulative effects of the intervention, which refers to the over-time change in the time series (as opposed to the short-time, single-day effect).

probability p is approaching significance ($p = .035$), the relative effect crosses the 95% confidence interval [−0.06%, 1.6%].[12]

With regards to insults (RQ1d), we find a small but statistically significant difference (posterior tail-area probability $p < .001$) between the cumulative predicted levels of insults in the counterfactual ($M = 22.88$, $SD = 0.11$) and the effects of the intervention ($M = 23.68$). The relative effect of the intervention was positive, indicating an increase in insults, and is larger compared to the relative effect of toxicity (3.51%); it is also within the bounds of the 95% confidence interval [2.5%, 4.5%].

With regards to profanity (RQ1e), we find a small, statistically significant, and negative difference (posterior tail-area probability $p = .007$) between the cumulative predicted levels of profanity in the counterfactual ($M = 24.39$, $SD = 0.12$) and the effects of the intervention ($M = 24.09$). Unlike prior results, the relative effects are negative (−1.2%), suggesting that the intervention decreased profanity relative to the counterfactual. This is also within the bounds of the 95% confidence interval [−2.2%, −0.25%].

With regards to threats (RQ1f), we find the largest (relative to our other results) statistically significant (posterior tail-area probability $p = .001$) difference between the average predicted levels of threats in the counterfactual ($M = 3.57$, $SD = 0.03$) and the effects of the intervention ($M = 3.83$). The relative effect of this intervention is 7.3%, indicating an increase in threats to LGBTQIA+ following Musk's acquisition, and is within the bounds of the 95% confidence interval [5.4%, 9.3%].

To explain these effects, we visualized the changes over time (Figure 1). Of the categories of aversive language, toxicity, insult, and threats permanently increased following Musk's acquisition of Twitter. In the case of identity attacks, we observe a short-term effect, suggesting that this effect was short-lived. With profanity, we actually see a decrease following Musk's acquisition, likely due to the ease of identifying this content using a small keyword list.

Our results provide mixed evidence for an overall increase in toxicity before and after Musk's acquisition of Twitter. As a post hoc analysis, we constructed GARCH models to understand the volatility of toxicity over time.

Post hoc analysis

Our analysis of RQ1 indicates that many aspects of aversive discourse increased following Musk's acquisition of Twitter. However, descriptives of the time series, as well as the variance in the cumulative effects of the intervention, also suggest an increase in volatility such that aversive discourse increased and decreased substantially following the intervention. In practice, this would suggest that there were large and haphazard fluctuations in how Musk attempted to address aversive discourse related to LGBTQIA+ discourse. While not an anticipated outcome at the onset of this study, the possibility of increasing volatility makes sense given the varied ways that Musk attempted to moderate discourse on Twitter (Paul and Dang, 2022; Tangalakis-Lippert, 2023).

To assess the volatility of aversive discourse after Musk's takeover, we utilized a GARCH model that assumes that the volatility of a time point is conditional on previous points in the time series. It is used to predict increasing or decreasing volatility. To estimate the volatility of each time series, we construct GARCH(1,1) models: the (1,1) indicates that we are estimating the volatility of time t based on the immediately prior time point, $t + 1$ (these are the most common models utilized in econometrics time series literature, see Hansen and Lunde, 2005). A GARCH(1,1) suggests increasing volatility if the GARCH term is statistically significant and positive.

First, we examine toxicity. The GARCH(1,1) model suggests a statistically significant GARCH term (coefficient = 0.99, $p < .001$), indicating that volatility increases following Musk's acquisition of Twitter. A model with the GARCH term, which helps predict volatility, is a better fit for modeling the time series than a model without the GARCH term ($\chi^2 = 14.51$, $p = .001$).

Next, we study volatility in the severe toxicity time series. The GARCH(1,1) model suggests a significant GARCH term (coefficient = 0.97, $p < .001$), indicating that the volatility of severe toxicity increases following Musk's acquisition. A model with the

Table 3. Results of negative binomial regression predicting engagement.

	Dependent variable: Engagement
Constant	3.54 (0.02)***
Toxicity	2.78 (0.08)***
Severe toxicity	0.26 (0.06)***
Insults	-4.32 (0.08)***
Time (Before Musk took over = Reference group)	0.08 (0.02)***
Interaction terms	
Time * Toxicity	-1.53 (0.11)***
Time * Severe toxicity	0.57 (0.09)***
Time * Insults	1.93 (0.10)***
	Log-likelihood −811,922.70
	theta 0.29*** (0.001)
	AIC 1,623,861.000

$N = 215{,}365$. The regression coefficient (β) is reported with the standard error in parentheses. ***$p < .001$.

GARCH term is a better fit for modeling the time series than a model without the GARCH term ($\chi^2 = 9.11$, $p = .002$).

We explore the volatility of identity attacks. The GARCH(1,1) model has a significant GARCH term (coefficient = 0.94, $p < .001$). The model with the GARCH term is a better fit for modeling the identity attacks time series than a model without the GARCH term ($\chi^2 = 5.83$, $p = .015$). This suggests that, even if there is no significant increase in identity attacks (as suggested by the intervention analysis), there is an increase in the volatility of identity attacks related to LGBTQIA+ discourse per day following Musk's acquisition of Twitter.

We examine insults. The GARCH(1,1) model has a significant and positive GARCH term (coefficient = 0.75, $p < .001$), indicating that volatility has increased following Musk's acquisition of Twitter. The model with the GARCH term is a better fit than a model without a GARCH term ($\chi^2 = 42.21$, $p < .001$).

After this, we turn to profanity. Similar to the other variables, we find a significant and positive GARCH term (coefficient = 0.98, $p < .001$), which suggests that the volatility of profanity discourse per day increases following Musk's acquisition. The model with the GARCH term is a better fit than a model without a GARCH term ($\chi^2 = 51.77$, $p < .001$).

Finally, we turn to threats. The GARCH(1,1) model has a significant and positive GARCH term (coefficient = 0.91, $p < .001$), suggesting that volatility increases following Musk's acquisition. The model with a GARCH term is a better fit than a model without a GARCH term ($\chi^2 = 79.42$, $p < .001$).

Cumulatively, these results indicate that, for all dimensions of aversive discourse as measured by the Google Perspective API, volatility increased following Musk's acquisition of Twitter.

RQ2 examined if there is a difference in the engagement level of tweets with toxicity after Musk took over, compared to before. First, we ran a stepwise regression for optimal feature selection. We selected the model based on the AIC, with lower values indicating

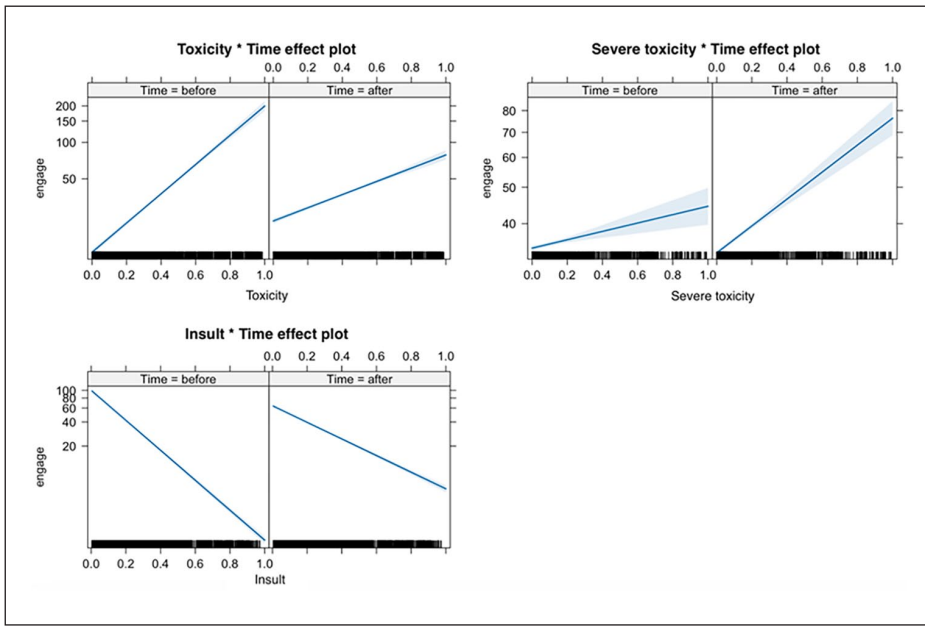

Figure 2. Effects of toxicity, severe toxicity, and insults, before and after Elon, on engagement. Shades around the lines indicate confidence intervals, and the rug plot shows the density of data points.

a better fit. Using forward-backward steps approach, the final model (AIC = 2,976,512) suggest toxicity (β = 68.99, $p = .015$), severe toxicity (β = 42.29, $p = .07$),[13] and insults (β = −120.56, $p < .001$) as significant predictors.

Based on this, we ran a negative binomial regression (see Table 3). We found that toxicity (β = 2.78, $SE = 0.08$, $p < .001$), severe toxicity (β = 0.26, $SE = 0.06$, $p < .001$), and time (β = 0.08, $SE = 0.02$, $p < .001$) are positive predictors of engagement, whereas insults negatively predict engagement (β = −4.32, $SE = 0.08$, $p < .001$). Our results show that people were more likely to engage with content that was toxic, severely toxic, or was posted after Musk took over Twitter, while content that contained insults was less likely to receive engagement.

The interactions between time and toxicity (β = −1.53, $p < .001$), time and severe toxicity (β = 0.57, $p < .001$), as well as time and insults (β = 1.93, $p < .001$), significantly predicted engagement. To explain the interaction terms, we present interaction plots (Figure 2).

The first plot, illustrating the interaction between *toxicity* and time, shows a decreasing slope. This suggests that post-Musk tweets with *toxicity* are less likely to engage Twitter users compared to the pre-Musk era. In contrast, the slope for *severe toxicity* increases, indicating that severely toxic content has gained more engagement after Musk's takeover. The plot for *insults* indicates an initial decrease in engagement levels, but this trend weakened post-Musk. That is, post-Musk, tweets with insults receive more engagement than before. In summary, our analyses suggest that after Musk's

acquisition of Twitter, there has been a noticeable change in engagement levels with aversive tweets. Tweets categorized as severely toxic and insulting have gained more engagement, while those considered toxic have experienced a decrease in engagement.

Discussion

Our aim with this study was to examine the larger societal problem of aversive content online by examining one cultural moment, when Elon Musk bought Twitter and dismantled many of the safety mechanisms intended to limit hate speech. By analyzing the discourse on the platform itself, we sought to understand whether Twitter became less hospitable to LGBTQIA+ people, a group Musk had repeatedly shown animosity toward (Yurcaba and Ingram, 2023), after he bought the platform. This study highlights the differences in Twitter discourse before and after Musk's takeover of Twitter in 2022, aiming to understand the shifts in the level of aversiveness in public conversations. We use the case of discourse about LGBTQIA+ people as an exemplar of the larger societal problem of online discord in the wake of Musk's changes to the Twitter platform.

Our time-series intervention analysis revealed that aversive language (toxicity, insults, profanity, and threats) about LGBTQIA+ people significantly changed after Musk's acquisition, although the differences were small. We observed an increase in toxicity, insults, and threats, but a decrease in profanity. This general rise in aversive content is consistent with research, which also noted a dramatic increase in hate speech following Musk's takeover of Twitter (Benton et al., 2022; Hickey et al., 2023). The finding regarding profanity decreasing deserves some unpacking. In some ways, it seems normatively positive that at least one type of aversive content decreased, but it is outside the scope of this study to untangle why that occurred. Future scholars should attempt to understand this more. One possibility is that profanity is relatively easy to spot and remove, and, indeed, one of the main types of content that gets deleted by moderators (Muddiman and Stroud, 2016; Vargo et al., 2024). Thus, even if Musk decreased moderation efforts, profanity would likely still be flagged as problematic and, likely, removed. However, we submit that it is normatively problematic that profanity decreased while other more divisive types of aversive content, such as threats, increased. In addition, we found that the volatility of all dimensions of aversive content (toxicity, severe toxicity, identity attack, insult, threats, profanity) increased, though the extent varied. This suggests that after Musk's takeover of Twitter, aversive content went up and down more frequently, compared to before. While some may find it heartening that we did not see dramatic increases in aversive content after Musk's takeover, even small increases can be sobering, given the overwhelming negative online climate. Furthermore, the volatility of the tweets raises concerns about instability, both for the platform and the users.[14] For the platform, this volatility may reflect constant changes in content moderation, which defeats the purpose of having consistent standards. For the user, the volatility of aversive language, due to its unpredictability, makes a platform feel inherently unsafe. Even if there was not much hate speech on one day, there is no guarantee that the user would not be exposed to aversive LGBTQIA+ posts the day after. This ambiguity can contribute to anxiety (Gu et al., 2020; Hoyt et al., 2022) and the sense that the platform is unsafe.

We argue that these findings warrant more study to understand if that is a quirk or a signal of greater escalation. Our findings shed light on at least one representation of the public conversation about LGBTQIA+ discourse. From the ritual communication viewpoint, our findings suggest people are talking about LGBTQIA+ people and issues in ways that are insulting and toxic and include identity attacks, threats, and profanity. While unsurprising, this offers early empirical evidence of the extent of online hate about LGBTQIA+ issues.

Analyzing the engagement with this aversive content reveals even more troubling trends. Our data show that severely toxic and insulting content, particularly tweets posted after Musk's takeover, received higher levels of engagement, while content that is merely toxic received less engagement. From a normative perspective, the increased engagement of Twitter users with aversive content poses a serious problem, especially since aversive language in social media posts can escalate the aversiveness of subsequent comments (Kim et al., 2021). This suggests a potential cycle where exposure to and interaction with aversive content may normalize and amplify inhospitality toward the minoritized group, further perpetuating more hostile online public discourse. This finding highlights that, without sufficient content moderation, social media can become a channel for spreading hate against marginalized groups, despite its potential as a platform for social justice and empowerment. Our research underscores the power of social media platforms to influence the public consciousness by changing corporate ownership or content moderation policies, and the troubling ramifications on society of those changes.

Conclusion

While our analysis spanned the 3 months before and after Musk's acquisition of Twitter and the immediate dismantling of much of the content moderation, the impact of updated content moderation policies and the dismissal of content moderators on online discussions may become more apparent over time. Our study serves as a preliminary observation, suggesting that a more detailed examination over a longer period would be beneficial. Although our focus was on LGBTQIA+ people, the inhospitality of online spaces likely affects other historically marginalized groups, including those defined by race or ethnicity. We acknowledge several limitations of this study. We focus on tweets in one language on one general topic, and we encourage researchers to conduct similar studies with a focus on other minoritized groups. In addition, our data provide no means to ascertain whether the tweet producers are from within or outside LGBTQIA+ communities, and we suggest that future studies consider this. We used machine-learning algorithms to measure content aversiveness; however, this approach presents potential challenges due to the context-dependent nature of the concept (Hede et al., 2021; Muddiman et al., 2019). Despite its limitations, the strength of the Perspective API lies in its ability to offer multiple dimensions of aversive content; we chose not to aggregate them to achieve a more contextualized understanding of aversiveness. The proprietary qualities of the Perspective API are a limitation because they make it hard to know exactly how the API defines different types of content, although, as described earlier in our paper, we provide conceptual definitions of each type. The discontinuation of Twitter API also makes it challenging to replicate our data, but this is a limitation that we have

to live with, given the dearth of reliable, alternative tools (Munger, 2019). Finally, a limitation of our time series modeling is that our data are averaged to the daily level. Future research can examine aversive language at more granular levels of temporality, such as at the hourly or minute level.

Acknowledgements

This research is a project of the Center for Media Engagement (CME), Moody College of Communication, at The University of Texas at Austin, United States. The authors thank E. Ciszek and Jenny Squibb for their assistance in compiling keywords for data collection; Shuting Yao and Lahari Siva Prasad Naraharisetty for their help with coding; and Natalie J. Stroud for her assistance with the project and the CME team for feedback on an earlier draft of this manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: It is part of the CME's connective democracy initiative, which is supported through a grant from the John S. and James L. Knight Foundation.

ORCID iDs

Gyo Hyun Koo https://orcid.org/0000-0002-3188-4588

Josephine Lukito https://orcid.org/0000-0002-0771-1070

Gina M Masullo https://orcid.org/0000-0002-4909-2116

Christian Staal Bruun Overgaard https://orcid.org/0000-0002-8774-0778

Notes

1. According to Elon Musk's Twitter account on 27 October 2022. https://x.com/elonmusk/status/1585841080431321088?lang=en
2. We refer to the platform as Twitter throughout because during the period when data were collected, Twitter had not yet been renamed X.
3. See: https://www.aclu.org/legislative-attacks-on-lgbtq-rights-2024
4. See https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US
5. We note that Twitter is not unique in contributing to disruptions in safe online conversations, as many social media platforms face similar challenges regarding transparency and accountability in platform governance (e.g. Thach et al., 2022).
6. The Online Supplemental Materials are available at: https://osf.io/eu7sp/?view_only=80e1ebf2eed64607bc64e82cb85319d7. Note that our list of keywords contains offensive terms. We also provide the number of tweets mentioning the keywords in Online Supplementary Materials S2.
7. To test the robustness of our findings from the 1% random sample (n=323,440), we conducted a stratified analysis. The full dataset (N=32,713,090) was divided into 35 batches, and every seventh batch was selected, yielding 3,972,260 tweets (~12.14%). Descriptive statistics and results are presented in Online Supplementary Materials S5. Overall, the findings are consistent with those from the 1% sample, with only minor differences observed.
8. To account for variation in the amount of posts day-by-day, we use a daily average for all time series analyses, including the aversive content metrics, rather than a cumulative amount.

9. This was done to compare treatment groups (posts created after Elon's takeover) with control groups (posts created before), comparing changes in outcomes over time between these two groups. This approach is effective for measuring the treatment effect on the treated and is typically implemented by coding time periods as 0 (time before the event) and 1 (time during and after the event) (Columbia University Mailman School of Public Health, n.d.; Torres-Reyna, 2015). For instance, researchers have used this approach to code posts created before and after COVID-19 to predict the presence of misinformation (Yussof et al., 2023), as well as before and during the pandemic to analyze trust in information sources (Bispo et al., 2023).
10. Toxicity = 7.25; Severe Toxicity = 2.21; Insult = 5.86. Research suggests using cutoff points of 10 for the VIF to identify severe multicollinearity (Bagheri and Midi, 2009).
11. The posterior tail-area probability *p*-value uses a posterior distribution (rather than a normal distribution in a typical *p*-value). While this value can be interpreted similarly to a *p*-value, it is preferable to rely on the confidence interval when the results are nearing significance. A posterior tail-area probability *p*-value is common to intervention analyses (e.g. Takyi and Bentum-Ennin, 2021).
12. When conducting an analysis using CausalImpact, researchers have used the confidence intervals (a more conservative estimate), rather than the posterior tail-area probability *p*-value, to assess statistical significance (e.g. Cervantes and Rambaud, 2020; Takyi and Bentum-Ennin, 2021).
13. We retain variables in the results with *p*-values above .05, focusing on model fit rather than the significance of individual predictors. Of note, when we ran a negative binomial regression without severe toxicity, it yielded consistent results: toxicity ($\beta = 2.33$, $p < .001$) and time ($\beta = .02$, $p = .01$) positively predicted engagement, while insult ($\beta = -3.37$, $p < .001$) negatively predicted it.
14. As noted in the methods section, GARCH models were originally developed to measure stock volatility. In market predictions, investors often struggle with the uncertainty of how to invest in highly volatile markets. Similarly, we anticipate that the increased volatility of aversive LGBTQIA+ content creates a more uncertain and therefore negative user experience.

References

Bispo JB, Douyon A, Ashad-Bishop K, et al. (2023) How trust in cancer information has changed in the era of COVID-19: patterns by race and ethnicity. *Journal of Health Communication* 28(3): 131–143.

Bagheri A and Midi H (2009) Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. *Journal of Mathematics and Statistics* 5(4): 311–321.

Barrie C (2023) Did the Musk takeover boost contentious actors on Twitter? *HKS Misinfo Review*. Available at: https://misinforeview.hks.harvard.edu/article/did-the-musk-takeover-boost-contentious-actors-on-twitter/

Barrie C and Ho JC (2021) academictwitteR: an R package to access the Twitter academic research product track v2 API endpoint. *Journal of Open Source Software* 6(62): 3272.

Benton B, Choi JA, Luo Y, et al. (2022) Hate speech spikes on Twitter after Elon Musk acquires platform. *School of Communication and Media Scholarship and Creative Works*, vol. 33. Available at: https://digitalcommons.montclair.edu/scom-facpubs/33

Borgogna NC, McDermott RC, Aita SL, et al. (2019) Anxiety and depression across gender and sexual minorities: implications for transgender, gender nonconforming, pansexual, demisexual, asexual, queer, and questioning individuals. *Psychology of Sexual Orientation and Gender Diversity* 6(1): 54–63. DOI: 10.1037/sgd0000306

Brock A (2012) From the Blackhand side: Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media* 56(2): 529–549.

Brock A (2018) Critical technocultural discourse analysis. *new media & society* 20(3): 1012–1030.

Brodersen KH, Gallusser F, Koehler J, et al. (2015) Inferring causal impact using Bayesian structural time series models. Available at: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-9/issue-1/Inferring-causal-impact-using-Bayesian-structural-timeseries-models/10.1214/14-AOAS788.full

Bronski M (2011) *A Queer History of the United States*. Boston, MA: Beacon Press.

Butler J (1999) *Gender Trouble: Feminism and the Subversion of Identity*. London: Routledge.

Carey J (1989) *Communication as Culture*. London: Routledge.

Cervantes PAM and Rambaud SC (2020) An empirical approach to the "Trump effect" on US financial markets with causal-impact Bayesian analysis. *Heliyon* 6(8): e04760.

Chen GM (2017) *Online Incivility and Public Debate: Nasty Talk*. London: Palgrave Macmillan.

Chen GM and Lu S (2017) Online political discourse: exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media* 61(1): 108–125.

Coe K, Kenski K and Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in website comments. *Journal of Communication* 64: 658–579.

Columbia University Mailman School of Public Health (n.d.) Difference-in-difference estimation. *Mailman School of Public Health, Columbia University*. Available at: https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation

Counts A and Nakano E (2023) July 19) Harmful content has surged on Twitter, keeping advertisers away. *Time*, 19 July. Available at: https://time.com/6295711/twitters-hate-content-advertisers/

Dang S (2022) Elon Musk manages free speech versus "hellscape" at Twitter. *Reuters*, 29 October. Available at: https://www.reuters.com/technology/elon-musk-takes-over-twitter-free-speech-limits-tested-2022-10-28/

Delgado R and Stefanic J (2012) *Critical Race Theory: An Introduction*. New York: New York University Press.

ElSherief M, Kulkarni V, Nguyen D, et al. (2018) Hate lingo: a target-based linguistic analysis of hate speech in social media. *Association for the Advancement of Artificial Intelligence*. Available at: https://arxiv.org/abs/1804.04257

Flores AR, Stotzer RL, Meyer I, et al. (2022) Hate crimes against LGBT people. National crime victimization survey, 2017-2019. *PLoS-ONE* 17: e0279363.

Foucault M (1985) *The History of Sexuality*. New York: Pantheon Books.

Francq C and Zakoian JM (2019) *GARCH Models: Structure, Statistical Inference and Financial Applications*. Hoboken, NJ: John Wiley & Sons.

Franses PH and Van Dijk D (1996) Forecasting stock market volatility using (non-linear) Garch models. *Journal of Forecasting* 15(3): 229–235.

Frimer JA, Aujla H, Feinberg M, et al. (2022) Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science* 14(2): 259–269.

Fung B (2021) January 9) Twitter bans President Trump permanently. *CNN*, 9 January. Available at: https://www.cnn.com/2021/01/08/tech/trump-twitter-ban/index.html

Gervais BT (2015) Incivility online: affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics* 12(2): 167–185.

GLAAD (n.d.) Glossary of Terms: LGBTQ. Available at: https://glaad.org/reference/terms (accessed 19 February 2024).

Gu Y, Gu S, Lei Y, et al. (2020) From uncertainty to anxiety: how uncertainty fuels anxiety in a process mediated by intolerance of uncertainty. *Neural Plasticity* 1: 8866386.

Guo M and Sun F (2020) Like, comment, or share? Exploring the effects of local television news Facebook posts on news engagement. *Journal of Broadcasting & Electronic Media* 64(5): 736–775.

Hall S (2000) Racist ideologies and the media. In: Marris P and Thornham S (eds) *Media Studies: A Reader*. New York: New York University Press, pp. 271–282.

Hansen PR and Lunde A (2005) A forecast comparison of volatility models: does anything beat a GARCH (1, 1)? *Journal of Applied Econometrics* 20(7): 873–889.

Harlow S (2015) Story-chatterers stirring up hate: racist discourse in reader comments on U.S. newspaper websites. *Howard Journal of Communications* 26(1): 21–42.

Hede A, Agarwal O, Lu L, et al. (2021) From toxicity in online comments to incivility in American news: proceed with caution. In: P*roceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Kyiv, 19–23 April, pp. 2620–2630. Available at: https://doi.org/10.18653/v1/2021.eacl-main.225

Heiss R, Schmuck D and Matthes J (2019) What drives interaction in political actors' Facebook posts? Profile and content predictors of user engagement and political actors' reactions. *Information, Communication, & Society* 22(10): 497–1513.

Hickey D, Schmitz M, Fessler D, et al. (2023) Auditing Elon Musk's impact on hate speech and bots. *Proceedings of the International AAAI Conference on Web and Social Media* 17(1): 1133–1137.

Hilbe JM (2011) *Negative Binomial Regression*. Cambridge: Cambridge University Press.

Hoyt DL, Hiserodt M, Gold AK, et al. (2022) Is ignorance bliss? Examining the effect of news media exposure on anxiety and depression during the COVID-19 pandemic. *The Journal of Nervous and Mental Disease* 210(2): 91–97.

Ingram D (2024) Elon Musk's transgender daughter, in first interview, says he berated her for being queer as a child. *NBC News*, 25 July. Available at: https://www.nbcnews.com/tech/tech-news/elon-musk-transgender-daughter-vivian-wilson-interview-rcna163665

Jikeli G and Soemer K (2022) Conversations about Jews on Twitter: recent developments since Elon Musk's takeover. *Computational and Mathematical Organization Theory*, 28 October. Available at: https://isca.indiana.edu/publication-research/social-media-project/Conversations-About-Jews-on-Twitter.-Recent-Developments-Since-the-Takeover-by-Elon-Musk.pdf

Karpf D (2018) Analytic activism and its limitations. *Social Media + Society* 4(1): 1–10.

Kim JW, Guess A, Nyhan B, et al. (2021) The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* 71(6): 922–946.

Klein A (2019) From twitter to charlottesville: analyzing the fighting words between the alt-right and antifa. *International Journal of Communication* 13: 297–318.

Kostić B and McGonagle T (2019) How social are new and social media for national minorities? Perspectives from the FCNM. *European Yearbook of Minority Issues Online* 16(1): 1–27.

Lundbergh S and Teräsvirta T (2002) Evaluating GARCH models. *Journal of Econometrics* 110(2): 417–435.

McGregor SC (2019) Social media as public opinion: how journalists use social media to represent public opinion. *Journalism* 20(8): 1070–1086.

Min S (2007) Online vs. face-to-face deliberation: effects on civic engagement. *Journal of Computer-mediated Communication* 12: 1369–1387.

Muddiman A and Stroud NJ (2016) 10 things we learned by analyzing 9 million comments from The New York Times. *Center for Media Engagement*. Available at: https://mediaengagement.org/research/10-things-we-learned-by-analyzing-9-million-comments-from-the-new-york-times/

Muddiman A and Stroud NJ (2017) News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication* 67: 586–609.

Muddiman A, McGregor SC and Stroud NJ (2019) (Re)claiming our expertise: parsing large text corpora with manually validated and organic dictionaries. *Political Communication* 36(2): 214–226.

Munger K (2019) The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society* 5(3): 1–4.

Nogara G, Pierri F, Cresci S, et al. (2023) Toxic bias: perspective API misreads German as more toxic. arXiv Preprint. Available at: https://arxiv.org/abs/2312.12651

Olteanu A, Castillo C, Boy J, et al. (2018) The effect of extremist violence on hateful speech online. In: *Proceedings of the International AAAI Conference on Web and Social Media*, Palo Alto, CA, 25–28 June. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/15040

Oz M, Zheng P and Chen GM (2018) Twitter versus Facebook: comparing incivility, impoliteness, and deliberative attributes. *New Media & Society* 20(9): 3400–3419. DOI: 10.1177/1461444817749516

Papacharissi Z (2004) Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New media & society* 6(2): 259–283.

Paul K and Dang S (2022) Exclusive: Twitter leans on automation to moderate content as harmful speech surges. *Reuters*, 3 December. Available at: https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/

Perrigo B (2023) A brief history of Elon Musk saying one thing and doing another at Twitter. *Time*, 27 April. Available at: https://time.com/6274774/elon-musk-twitter-u-turns/

Poole E, Giraud EH and de Quincey E (2021) Tactical interventions in online hate speech: the case of #stopIslam. *new media & society* 23(6): 1415–1442.

Popli N (2022) As Elon Musk buys Twitter, the right is celebrating. *Time*, 28 October. Available at: https://time.com/6226238/twitter-elon-musk-right-wing-influencers-politicians-celebrate/

Pozzobon L, Ermis B, Lewis P, et al. (2023) On the challenges of using black-box APIs for toxicity evaluation in research. arXiv Preprint. Available at: https://arxiv.org/abs/2304.12397

Ribé MM, Kaltenbrunner A and Keefer JM (2021) Bridging LGBT+ content gaps across Wikipedia language editions. *The International Journal of Information, Diversity, & Inclusion* 5(4): 90–131.

Riedl MJ, Masullo GM and Whipple KN (2020) The downsides of digital labor: exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior* 107: 106262.

Ripley B, Venables B, Bates DM, et al. (2023) Package "MASS." *CRAN R* 538: 113–120.

Rogers R (2020) Deplatforming: following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35(3): 213–229.

Rohlinger DA, Rose K, Warren S, et al. (2023) Does the Musk Twitter takeover matter? Political influencers, their arguments, and the quality of information they share. *Socius: Sociological Research for a Dynamic World* 9: 1–15.

Salisu AA and Gupta R (2021) Oil shocks and stock market volatility of the BRICS: a GARCH-MIDAS approach. *Global Finance Journal* 48: 100546.

Salminen J, Sengün S, Corporan J, et al. (2020) Topic-driven toxicity: exploring the relationship between online toxicity and news topics. *PLOS ONE* 15(2): e0228723. DOI: 10.1371/journal.pone.0228723

Sobieraj S (2020) *Credible Threat: Attacks against Women Online and the Future of Democracy*. Oxford: Oxford University Press.

Takyi PO and Bentum-Ennin I (2021) The impact of COVID-19 on stock market performance in Africa: a Bayesian structural time series approach. *Journal of Economics and Business* 115: 105968.

Tangalakis-Lippert K (2023) A new California law regulating online content moderation will force X to reveal how it reviews hate speech. *Business Insider*, 29 December. Available at: https://www.businessinsider.com/elon-musk-x-content-moderation-polices-california-law-ab587-2023-12

Tawa K and Bunts W (2022) How queer visibility threatens white power. *The Center for Law and Social Policy*. Available at: https://www.clasp.org/blog/how-queer-visibility-threatens-white-power/

Thach H, Mayworm S, Delmonaco D, et al. (2022) (In)visible moderation: a digital ethnography of marginalized users and content moderation on Twitch and Reddit. *new media & society* 26(7): 5839.

Torres-Reyna O (2015) Difference-in-Differences (DID), August. Available at: http://www.princeton.edu/~otorres/

United Nations (2021) Report: online hate increasing against minorities, says expert. *United Nations Human Rights Report*, 23 March. Available at: https://www.ohchr.org/en/stories/2021/03/report-online-hate-increasing-against-minorities-says-expert

Vargo C, Masullo GM and Hopp T (2024) Deciding to delete posts on Reddit: what factors influence content removal. In: *Proceedings of the Digital and Social Media Track of the 57th Hawaii International Conference on System Sciences*, pp. 2593–2602. Available at: https://hdl.handle.net/10125/106696

Vargo CJ and Hopp T (2020) Fear, anger, and political advertisement engagement: a computational case study of Russian-linked Facebook and Instagram content. *Journalism & Mass Communication Quarterly* 97(3): 743–761.

Votta F (2021) peRspective: interface to the "Perspective" API (Version 0.1.1). Available at: https://github.com/favstats/peRspective

Wang R, Zhang Y, Suk J, et al. (2024) Empowered or constrained in platform governance? An analysis of Twitter users' responses to Elon Musk's takeover. *Social Media + Society* 10(3): 1–16.

Weber M, Huber M, Auch M, et al. (2025) Digital guardians: can GPT-4, perspective API, and moderation API reliably detect hate speech in reader comments of German online newspapers? arXiv Preprint. Available at: https://arxiv.org/abs/2501.01256

Wikipedia contributors (n.d.) Category: homophobic slurs. *Wikipedia The Free Encyclopedia*. Available at: https://en.wikipedia.org/wiki/Category:Homophobic_slurs (accessed 7 February 2024).

Wile R (2022) A timeline of Elon Musk's takeover of Twitter. *NBC News*, 17 November. Available at: https://www.nbcnews.com/business/business-news/twitter-elon-musk-timeline-what-happened-so-far-rcna57532

XDevelopers (2023) 'Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead.', X (formerly Twitter), 2 February. Available at: https://x.com/XDevelopers/status/1621026986784337922 (accessed 4 July 2025).

Yurcaba J and Ingram D (2023) A year after Elon Musk bought Twitter, LGBTA people say it has become toxic. *NBC News*, 27 October. Available at: https://www.nbcnews.com/nbc-out/out-news/year-elon-musk-bought-twitter-lgbtq-people-say-become-toxic-rcna122154

Yussof I, Ab Muin NF, Mohd M, et al. (2023) Breast cancer prevention and treatment misinformation on Twitter: an analysis of two languages. *Digital Health* 9: 10.1177/20552076231205742.

Author biographies

Gyo Hyun Koo (PhD, The University of Texas at Austin) (she/her) is an Assistant Professor in the Department of Communication, Culture and Media Studies at Howard University. Her research focuses on the role of communication technology in shaping how people perceive, process, and interact with online information and news.

Josephine Lukito (she/her) is an Assistant Professor at the University of Texas at Austin's School of Journalism and Media and the Director of the Media & Democracy Data Cooperative. Jo's work uses computational and machine learning approaches to study political language, with a focus on harmful digital content and multi-platform discursive flows. She also studies data access for researchers and journalists. Her work has been published in top-tiered journals such as *Political Communication* and *Social Media + Society*, and featured in *Wired*, *Columbia Journalism Review*, and *Reuters*.

Gina M Masullo (PhD, Syracuse University) (they/them) is Associate Director of the Center for Media Engagement and an Associate Professor in the School of Journalism and Media, both at The University of Texas at Austin, United States. Their research focuses on how the digital space both connects and divides people and how that influences society, individuals, and journalism. They are the author of *Online Incivility and Public Debate: Nasty Talk* and *The New Town Hall: Why We Engage Personally with Politicians* and co-editor of *Scandal in a Digital Age*. Their latest book, *Midlife Sapphic Revelation in the Digital Age: How Digital Media Support Coming Out Late*, will be published in 2025.

Christian Staal Bruun Overgaard (MSc, University of Southern Denmark) is a Knight Research Associate at the Center for Media Engagement and a doctoral student at the School of Journalism and Media at The University of Texas at Austin. His research interests include news, social media, and political polarization.

Bek Orr (PhD, Syracuse University) is an Associate Professor of Women and Gender Studies and Sociology at State University of New York, Brockport, USA. Their areas of research include queer communities, histories, archives, and digital communities. Their recent publications have appeared in *Feminist Media Studies*, *Journal of Fat Studies*, and *Journal of Feminist Scholarship*.