Exposure to Marginally Abusive Content on Twitter

Jack Bandy[1*], **Tomo Lazovich** [2*]

[1] Northwestern University
[2] Northeastern University
jackbandy@u.northwestern.edu, t.lazovich@northeastern.edu

Abstract

Social media platforms can help people find connection and entertainment, but they can also show potentially abusive content such as insults and targeted cursing. While platforms do remove some abusive content for rule violation, some is considered "margin content" that does not violate any rules and thus stays on the platform. This paper presents a focused analysis of exposure to such content on Twitter, asking (RQ1) how exposure to marginally abusive content varies across Twitter users, and (RQ2) how algorithmically-ranked timelines impact exposure to marginally abusive content. Based on one month of impression data from November 2021, descriptive analyses (RQ1) show significant variation in exposure, with more active users experiencing higher rates and higher volumes of marginal impressions. Experimental analyses (RQ2) show that users with algorithmically-ranked timelines experience slightly lower rates of marginal impressions. However, they tend to register more total impression activity and thus experience a higher cumulative volume of marginal impressions. The paper concludes by discussing implications of the observed concentration, the multifaceted impact of algorithmically-ranked timelines, and potential directions for future work.

Introduction

As digital media platforms continue to reshape public dialogue and information flows, urgent challenges have emerged regarding exposure to various types of harmful content. In a 2020 Pew Research survey, 64% of U.S. adults said that social media had a "mostly negative" effect on the public, citing misinformation and hateful content as the two most common reasons (Auxier 2020). These and other types of harmful content pose significant threats to well-engaged, well-informed publics, and researchers, journalists, and platforms have spent significant resources to address them.

New questions have also emerged regarding more nuanced categories of potentially harmful "margin content:" content that does not violate platform rules around abusive and hateful conduct, but bears abusive characteristics and may still be harmful to some users. For example, a post may contain insulting language that does not violate platform rules regarding hate speech or harassment. Platforms would not remove such a post; however, many users may consider it harmful. This is true for various types of margin content, and particularly for marginally abusive content such as insults or targeted cursing.[1]

This paper aims to address some foundational research questions regarding exposure to marginal content on Twitter. Namely, (RQ1) how does exposure to marginally abusive content vary across users? And, (RQ2) how do algorithmic timelines (intended to elevate safe and relevant content) affect exposure to marginally abusive content? The paper offers the first large-scale empirical findings of this nature, using real-world impression data from hundreds of millions of users. In particular, this paper's contributions include:

- Descriptive evidence that exposure to marginally abusive content varies significantly across Twitter users, both in terms of rates and volumes
- Experimental evidence that users with algorithmically-ranked timelines experience slightly lower marginal impression rates, but a higher cumulative volume of marginal impressions due to increased activity

Related Work

Here we review two broad research areas that inform our work. The first area relates to marginal content on social media platforms and includes descriptive studies, qualitative work, and experiments testing potential interventions. The second area is essentially the field of algorithm auditing, and we focus specifically on audit studies that explore how algorithmic systems impact social media platforms.

Marginal and Abusive Content on Platforms

Social media platforms have brought waves of challenges and research questions when it comes to harmful content online. Early research in this area focused on content that directly broke platform rules, especially human labor involved in reviewing and moderating such content (Roberts 2019; Gillespie 2018; Gray and Suri 2019). However, there is a large gray area between healthy, permissible content and violative content that platforms remove. Content in this gray area varies in type as well as in severity (Scheuerman et al.

[*] Work was performed while authors were employed by Twitter

[1] For simplicity, some places in the text use "margin" or "marginal" as stand-ins for "marginally abusive."

2021), and is sometimes referred to as "borderline content" or "margin content." Some platforms have explicitly stated that they aim to reduce exposure to margin content but will not remove it from the platform (Alexander 2019; Constine 2018).

While some related work explores margin content from the perspective of content creators (Caplan and Gillespie 2020), our work is one of the first to characterize *exposure* to margin content for general users. Also, margin content is a fairly broad category that can include graphic violence, adult content, content about regulated goods, and other subcategories. For this paper, we focus on marginally abusive content such as insults and targeted cursing.

Many different studies have contributed to an understanding of abusive content online. This includes some large-scale quantitative work analyzing different types of bullying in games (Kwak, Blackburn, and Han 2015) and other online contexts (Bellmore et al. 2015). Mozilla Foundation has published a study of "regretful views" on YouTube, based on 37,380 users who reported seeing different types of problematic videos (Mozilla Foundation 2021).

Some qualitative work has explored users' day-to-day experiences with respect to online safety (Redmiles, Bodford, and Blackwell 2019) and their strategies for dealing with abusive content (Vitak et al. 2017). Finally, some research has tested potential interventions, such as enhanced processes for reviewing and reporting harassment (Matias et al. 2015). One study tested a mechanism that prompted users before posting potentially offensive content to Twitter, finding that 34% of users revised their Tweet or did not send it at all (Katsaros, Yang, and Fratamico 2021), thus reducing the supply of potentially abusive content.

A recurring theme from this prior research is that some groups of users are disproportionately affected by exposure to abusive content. For example, studies suggest that women face particularly severe harassment and toxic behavior in online spaces (Amnesty International 2018; Vogels 2021). Researchers point out that this may be associated with the gendered nature of content moderation (Nurik 2019), which often pressures marginalized groups into conformity with broader social normativities (Feuston, Taylor, and Piper 2020).

The disproportionate impact of abusive content is a guiding point for our work. Despite the aforementioned initial evidence and hypotheses, there has been limited work to empirically characterize how user experience varies in terms of exposure to marginally abusive content. While our analysis does not explore specific demographic groups, it does analyze general variation and concentration patterns in exposure to marginal content. Specifically, RQ1 explores how exposure to marginally abusive content varies across Twitter users, based on large-scale empirical evidence.

Evaluating Algorithmic Impact on Platforms

Another key research area related to our work focuses on the real-world impact of algorithmic systems. While algorithm auditing is a fairly broad area with applications in finance, news, hiring, policing, commerce, and more, our work is most closely related to research evaluating algorithmic recommendation systems used by social media platforms.

One of the first and largest evaluations of social media algorithms analyzed the Facebook News Feed using a sample of users who self-reported their ideological affiliation (Bakshy, Messing, and Adamic 2015). The study focused on exposure to partisan news and opinion, and found that the News Feed algorithm only slightly exacerbated a partisan echo chamber effect. While some research has echoed these findings (Bechmann and Nielbo 2018; Bruns 2019), additional research has expanded the scope to other platforms, other types of content, and other effects besides echo chambers.

A growing body of research focuses specifically on effects of Twitter's algorithmically-ranked timeline. Many studies explore its impact on the dissemination of news media, as one review points out (Orellana-Rodriguez and Keane 2018), with the spread of false or misleading news as a common focal point (Grinberg et al. 2019). Recent work has also emphasized partisan variation and source reliability. A recent study showed that algorithmic timelines disproportionately amplified content from right-leaning politicians in six out of seven countries (Huszár et al. 2022). In terms of source reliability, one study found through agent-based testing that Twitter's algorithmic timeline may slightly benefit "junk news" websites (Bandy and Diakopoulos 2021). Corroborating these findings, *The Economist* found that algorithmic timelines amplified less reliable sources more than reliable sources (The Economist 2021).

Our work joins some other early research efforts at the intersection of marginal content and algorithmic impact. One recent experiment showed that exposure to higher rates of hostile or "outrage" content makes users more likely to share similar outrage content themselves (Brady et al. 2021). This finding is consistent with prior work showing that content creation is a social and collective process, whether crafting Wikipedia articles (Nagar 2012) or spreading strategic disinformation (Starbird, Arif, and Wilson 2019). In short, users adopt communication norms they perceive from other users, and algorithmic systems such as the Twitter timeline can play a role in establishing those norms.

Our work builds on prior research in several ways. Broadly, we aim to characterize exposure to marginally abusive content on Twitter, recognizing that content exposure plays an important role in establishing communication norms on the platform (Brady et al. 2021). We also leverage the same large-scale experiment used in recent work (Huszár et al. 2022; The Economist 2021), in which a random sample of global Twitter users are assigned to only see reverse-chronological timelines. As a point of distinction, rather than analyzing exposure to political content or news websites, we analyze how algorithmically-ranked timelines affect exposure to marginally abusive content (RQ2). In short, related work guides us to the following research questions:

- (**RQ1**) How does exposure to marginally abusive content vary across Twitter users?
- (**RQ2**) How do algorithmic timelines affect exposure to marginally abusive content?

Category	Example Tweet
Advocates for consequences to livelihood	“Make sure she loses her job”
Targeted cursing	“You can go to hell!”
Claims of mental inferiority	“Why do idiots get to vote?”
Claims of moral inferiority	“Those people are pedophiles”
Other insults	“I hate those people”

Table 1: Examples of the categories of potential marginally abusive content considered for this research study.

Methods

To address the two main research questions regarding variation (RQ1) and algorithmic impact (RQ2) of exposure to marginally abusive content on Twitter, this paper uses descriptive and experimental methods. While we do not introduce any novel methods, this section describes relevant datasets, metrics, and other methodological details, starting with an operationalization of marginally abusive content.

Classifying Marginally Abusive Content

Identifying and classifying abusive content on social media platforms is a difficult process. For one, abusiveness can be subjective — a piece of content may offend some people but not others. Additionally, the massive scale of social media platforms makes it prohibitively costly to manually review every piece of content to determine whether it is abusive, marginally abusive, or permissible. Thus only a small portion of Tweets can be manually labeled, potentially introducing selection biases.

For the purposes of this paper, we consider several categories of potential marginally abusive content. These categories are based on commonly used concepts of “toxic content” (Davidson et al. 2017; Chandrasekharan et al. 2017). Tweets are considered “toxic” if they fall into one of several different categories, including those shown in Table 1.

To identify marginally abusive Tweets at scale, we relied on a machine classifier which has been developed in part based on the categories in Table 1. It is based on a pre-trained BERT model (Devlin et al. 2018) and was fine-tuned on a corpus of human-annotated Tweets. We provide three points of validation for this classifier, based on (1) previous work, (2) a subset validation corpus from November 2021, and (3) a disjoint validation corpus from February 2022.

First, previous work has used the same BERT-based model to help identify offensive content during Tweet composition (Katsaros, Yang, and Fratamico 2021). This work used the model as part of an algorithm to identify offensive reply Tweets. In a sample of 1,929 Tweets which the algorithm classified as offensive, 95% were labeled marginally abusive by human annotators. When the model was used to prompt users before posting potentially toxic Tweets, 34% of users revised their Tweet or did not send it at all (Butler and Parella 2021). In other words, both annotators and end users indicate the model captures potentially abusive content, although this validation only applies to reply Tweets.

We thus also validated the classifier using two sets of Tweets, the first being a corpus of 65,711 Tweets that received impressions in November 2021 (i.e. a subset of Tweets from the main datasets). These were randomly sampled with a weighting based on impressions, and each was annotated by five English-speaking individuals who were trained for the task. The annotation task is a series of questions used to identify the type and target(s) of the Tweet (as in Table 1), and a decision matrix determines the final label. Our classifier achieved an overall accuracy of 98% on this corpus. Furthermore, 75% of the Tweets identified as marginally abusive by the classifier were also identified as “toxic” by at least one annotator, indicating fairly strong precision considering the scale and complexity of the task.

As a third and final point of validation, we used a corpus of 56,003 Tweets that received impressions in February 2022 — a disjoint set of Tweets from the main datasets analyzed in the paper. As with the subset validation corpus, each Tweet was annotated by five English-speaking individuals. Inter-annotator agreement was strong, with 96% of toxic Tweets receiving majority votes from annotators. Among all Tweets determined to be toxic:

- 12% were labeled toxic by all 5 annotators
- 29% were labeled toxic by 4 annotators
- 55% were labeled toxic by 3 annotators
- 3% were labeled toxic by 2 annotators[2]

The classifier achieved fairly strong performance on the February 2022 corpus, with 98% overall accuracy. The classifier’s overall precision on the corpus was 52%, and overall recall was 25%. We discuss the implications of this performance in the limitations section. Similar to the November 2021 corpus, 77% of the Tweets identified as marginally abusive by the classifier were also identified as “toxic” by at least one annotator, which signals strong construct validity for the purposes of our analysis.

Data

This study uses real-world exposure data collected internally at Twitter — specifically, Tweet impression data. An *impression* is registered when at least 50% of a Tweet is visible in a user’s timeline for at least 0.5 seconds.

We use two datasets, one for RQ1 and one for RQ2. Both datasets span November 2021 and are not limited to any specific geographic region. As detailed below, the RQ1 dataset is more inclusive for the descriptive analysis, while the dataset for RQ2 uses an ongoing, randomized experiment described in prior work (Huszár et al. 2022), and represents a sample of the accounts in RQ1 dataset.

Dataset for RQ1 To address variation in exposure to marginally abusive content (RQ1), we used a dataset of all impressions on English-language Tweets for the full month

[2] Due to rounding, these proportions do not sum to 100%

	Accounts	Impressions
RQ1 Dataset	200M+	188B
RQ2 Dataset	630k reverse-chronological	842M
	13M algorithmically-ranked	24B

Table 2: Summary details for both datasets analyzed in the study. Both datasets span the full month of November 2021 and are not restricted to any specific geographic region.

of November 2021. This dataset includes every global Twitter account with at least one impression on an English-language Tweet. In total, this included over 200 million accounts and 188 billion impressions on English tweets. This includes accounts and impressions that also appear in the RQ2 dataset.

Dataset for RQ2 The second research question asks how algorithmic timelines impact exposure to marginally abusive content. To address this question, we leverage data from a long-term Twitter experiment which randomly excludes some accounts from the algorithmically-ranked timelines (introduced in 2016). At the time we collected our dataset, the randomized experiment included 630,000 accounts in the reverse-chronological timeline group and 13 million accounts in the algorithmically-ranked timeline group. Most Twitter accounts are not included in the experiment.

Prior work has used the same "holdback" experiment to measure algorithmic amplification of content from politicians and news organizations (Huszár et al. 2022). Here we provide the most relevant details of the experiment.

Broadly, users in the reverse-chronological group only experience timelines sorted in reverse-chronological order, while users in the algorithmically-ranked group experience the same timeline as most global Twitter users. Algorithmic timelines select, filter, and rank Tweets using a number of different systems, sometimes adding "recommended" Tweets from out-of-network accounts the user does not follow. Among the systems used for this process are machine learning models that predict engagement, content relevance, and more. Notably, some algorithms that power ranked timelines are intended to reduce exposure to marginally abusive content, especially from out-of-network accounts. These algorithms do not impact the reverse-chronological timeline, which only shows in-network content (Tweets and Retweets from followed accounts).

Some services shape content exposure in both reverse-chronological and algorithmically-ranked timelines. This includes promoted Tweets from advertisers (omitted from this analysis), Tweets that are hidden or displayed with a warning label due to platform rules, and Tweets that are hidden because a user blocked the Tweet's author or muted specific terms in the Tweet. These services are available to all users.

Finally, users in the algorithmically-ranked group do have the option to turn off algorithmic ranking and use a reverse-chronological timeline, though this is rare. The experiment thus runs at the account level rather than the session level.

Metrics

For both RQ1 and RQ2, we use a number of metrics to analyze marginal impressions in terms of volume, rate, and variation. In addition to simple distributional samples and measures of central tendency, we calculate concentration metrics (Farris 2010) including the Gini coefficient (Lorenz 1905; Gini 1912) and the top 1% share, described below:

- **Gini coefficient**: ranges from 0 to 1, with 0 indicating perfect equality and 1 indicating maximum inequality (e.g. one user accounting for 100% of impressions)
- **Top 1% share**: the proportion of impressions that come from the top 1% of users, ranges from 0% to 100%

Another key metric for this study is the *marginal impression rate*, which we calculate for each account based on English-language Tweets. Marginally abusive Tweets were identified using the BERT-based classifier described at the beginning of the Methods section. The marginal impression rate was calculated as the portion of total impressions on English-language Tweets that were classified as marginally abusive Tweets — that is, marginal English impressions divided by total English impressions).

One key shortcoming of the marginal impression rate metric is that it does not capture the overall timeline experience for users who see Tweets in multiple languages. We discuss this further in the limitations section.

Results

Overall, results show that exposure to marginally abusive Tweets is highly varied and distinctively concentrated, particularly among active accounts (RQ1). Our analysis also shows that algorithmically-ranked timelines slightly reduce marginal impression *rates* (RQ2). However, accounts with algorithmically-ranked timelines tend to register more total impression activity, which results in a higher cumulative *volume* of marginally abusive impressions.

Variation in Exposure (RQ1)

Descriptive analyses show significant variation in the volume of exposure to marginally abusive Tweets. Among accounts with at least one impression in November 2021, 71% registered no marginal impressions, and 87% registered fewer than five. However, accounts in the 99th percentile (in terms of marginal impression volume) saw 142+ marginally abusive Tweets, indicating that marginal exposures are highly concentrated among a small number of accounts.

Total impressions were also highly concentrated in our data. The median account registered 32 impressions on English-language Tweets over the course of November 2021 (about one impression per day), while accounts in the 99th percentile registered 13,376+ (445+ per day). Table 3 demonstrates this variation through distribution samples for marginal English impressions and total English impressions.

By several measures, marginal Tweet impressions are even more concentrated than total English Tweet impressions. The Gini coefficient was 0.88 for overall impressions, but 0.94 for marginally abusive impressions (see Lorenz

	1%	25%	50%	75%	99%
Total Imps.	1	5	32	247	13,376
Daily Imps.	0	0	1	8	445
Marginal Imps.	0	0	0	1	142
Daily Marginal Imps.	0	0	0	0	4

Table 3: Total and average daily volume of impressions on English Tweets and marginally abusive English Tweets, collected from all Twitter accounts in November 2021 (i.e. RQ1 dataset). Distribution samples (at the 1st, 25th, 50th, 75th and 99th percentiles) show the volume of exposure to marginally abusive content varies considerably across different users. These statistics only count impressions on English Tweets and thus do not reflect overall user experiences.

curves in Figure 1). Similarly, the top 1% share was 28% for overall impressions, and 45% for marginally abusive impressions (i.e. the top 1% of accounts registered 45% of all marginal impressions). In the discussion section, we note that this high concentration has mixed implications.

We also find variation in the *rate* of marginally abusive impressions. Because most accounts registered no marginal impressions in our dataset, most also experienced a marginal impression rate of 0.0%. However, the marginal impression rate at the 99th percentile was 6.3%, meaning these accounts experienced approximately 1 marginally abusive impression for every 16 total impressions.

Given the heavy skew and variation, we also calculated the marginally abusive impression rate for different groups of accounts based on overall activity. Specifically, we used five logarithmic bins based on average daily impressions on English Tweets. During November 2021,

- 49.4% of Twitter accounts averaged less than 1 English Tweet impression per day
- 27.5% averaged 1-10 per day
- 16.7% averaged 10-100 per day
- 6.2% averaged 100-1,000 per day
- 0.1% averaged more than 1,000 per day

As shown in Figure 2, groups with more total impressions per day averaged higher marginal impression rates. For example, accounts averaging 100-1,000 total impressions per day experienced a mean marginal impression rate more than twice as high as accounts averaging less than one impression per day (0.9% vs. 0.4%).

Notably, the different rates in Figure 2 imply a nonlinear relationship between total impressions and marginally abusive impressions — as accounts register more total impressions, they are more likely to register marginal impressions.

Algorithmic Impact (RQ2)

Experimental analysis shows that algorithmic timelines have multifaceted effects. Compared to a random sample of 630k accounts that only see reverse-chronological timelines, a sample of 13M accounts with algorithmically-ranked timelines experienced slightly lower rates of marginally abusive impressions. However, as detailed below, accounts with algorithmic timelines experienced a higher cumulative volume of marginal impressions due to an increase in total activity.

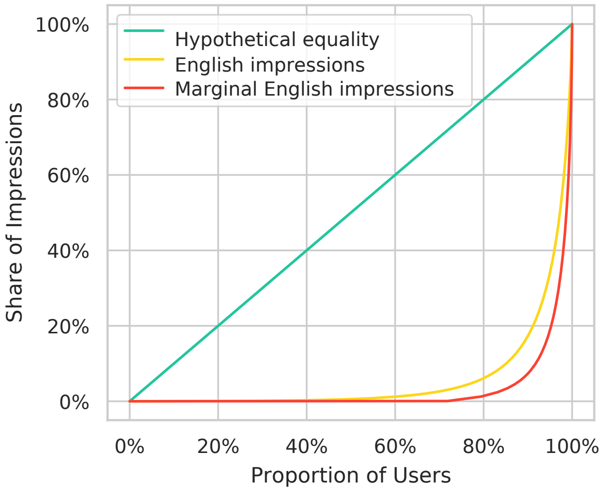

Figure 1: Lorenz curves representing impressions on English-language Tweets and marginally abusive English-language Tweets, with marginally abusive impressions being more concentrated. The Gini coefficient is 0.88 for the English impressions curve and 0.94 for the marginal impressions curve.

In terms of rates, the overall mean marginal impression rate was 0.55% for accounts with algorithmically-ranked timelines, and 0.59% for accounts with reverse-chronological timeline. These aggregate statistics obfuscate some effects: as shown in Figure 3, the reduction effect is more substantial for accounts with moderate levels of impression activity (10-100 or 100-1,000 impressions per day). However, the effect is insignificant for accounts with lower activity levels.

Despite lower rates of marginally abusive impressions, algorithmic timelines increase overall activity, resulting in a greater volume of total impressions as well as a greater cumulative volume of marginal impressions. Table 4 includes distribution samples to illustrate the increased activity resulting from algorithmic timelines. The median account(s) in the reverse-chronological timeline group recorded 95 impressions on English tweets in November 2021, while the median account(s) in the algorithmically-ranked timeline group recorded 116 — a median increase of 21 impressions over the course of the month.

Figure 4 shows the mean increase in impression activity: accounts with algorithmic timelines averaged roughly 500 extra impressions per account over the course of November 2021, compared to accounts with reverse-chronological timelines. As noted elsewhere, the distribution of impressions is extremely skewed, so this metric should be interpreted carefully.

The increased engagement from algorithmic timelines has important implications for exposure to marginally abusive

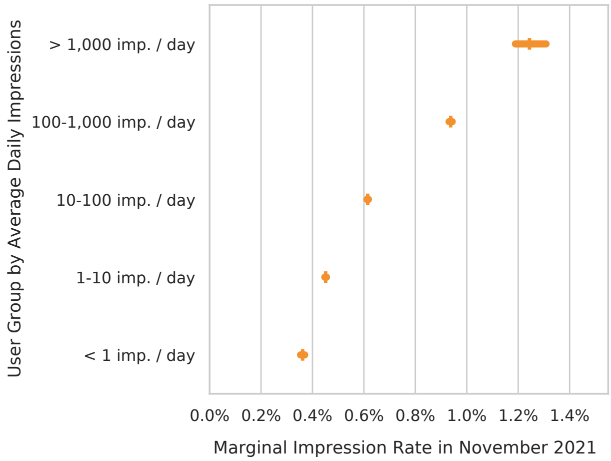

Figure 2: Marginally abusive impression rates for different groups of accounts, based on average English impressions per day in November 2021. More active accounts experienced higher rates of marginally abusive content. Vertical lines represent point estimates and horizontal bars represent bootstrapped 99% confidence intervals (1,000 iterations).

		1%	25%	50%	75%	99%
Total	Reverse-Chron.	1	16	95	520	24,174
	Algo.-Ranked	1	18	116	696	31,905
Marginal	Reverse-Chron.	0	0	0	3	309
	Algo.-Ranked	0	0	0	3	391

Table 4: Total impressions on English Tweets and on marginally abusive English Tweets, collected in November 2021 for accounts in the timelines quality holdback experiment. Accounts with algorithmically-ranked timelines tend to register a higher volume of total impressions.

content. Despite a lower marginal impression *rate*, accounts with algorithmic timelines averaged a greater cumulative volume of marginally abusive impressions compared to accounts with reverse-chronological timelines. The distribution is highly skewed (see Table 4), but the effect amounts to a mean increase of nearly four marginally abusive impressions over the course of the month — from approximately 15 per account for reverse-chronological timelines to 19 for algorithmic timelines, as shown in Figure 5.

In summary, results show that algorithmically-ranked timelines have a multifaceted effect on exposure to marginally abusive content. On one hand, they lower the rate of marginal impressions compared to reverse-chronological timelines. But they also tend to increase total impression activity, such that algorithmic timelines end up generating a higher cumulative volume of marginally abusive impressions per account.

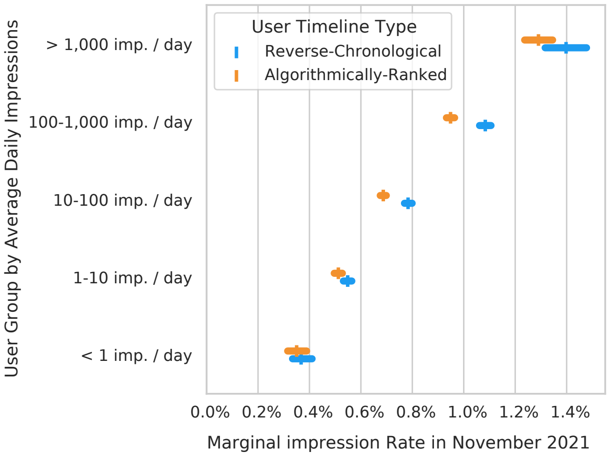

Figure 3: Accounts with algorithmically-ranked timelines experienced lower rates of marginally abusive impressions, though the effect size varies across groups of accounts with different activity levels. Vertical lines represent point estimates and horizontal bars represent bootstrapped 99% confidence intervals (1,000 iterations).

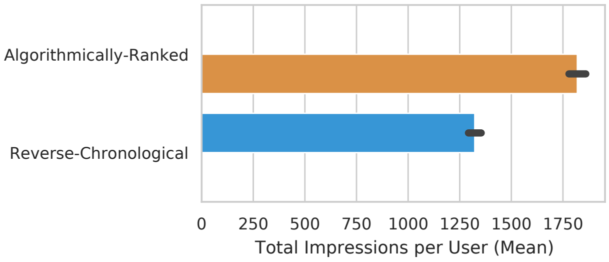

Figure 4: Accounts with algorithmically-ranked timelines registered a greater volume of total impressions, as shown here by the difference of means. The distribution is highly skewed, see Table 4 for distribution samples. Bars represent bootstrapped 99% confidence intervals (1,000 iterations).

Discussion

Here we interpret results from our analysis, discussing the observed concentration in exposure to marginally abusive content, the multifaceted impact of algorithmically-ranked timelines, and potential directions for future work. Before discussing these points, it is important to note some key limitations, as well as ethical considerations related to our study.

Limitations

Our work is subject to a number of limitations, mostly related to construct validity. First and foremost, as noted in the methods section, identifying marginally abusive content is a complex task given (1) the subjective nature of abusiveness and (2) the large scale of social media platforms. While the classifier used in this work was sufficient for addressing our research questions, its performance only allows us to capture

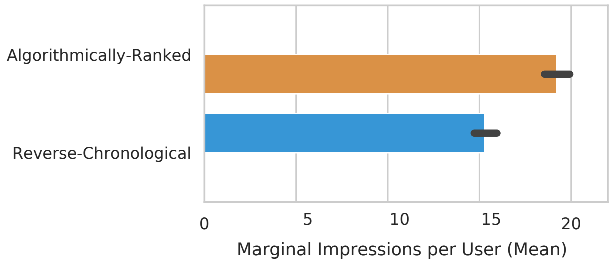

Figure 5: Despite lower marginal impression rates, accounts with algorithmically-ranked timelines registered a greater volume of marginally abusive impressions due to higher total impressions activity. Bars represent bootstrapped 99% confidence intervals (1,000 iterations).

approximate exposure to marginally abusive content. Given the low recall of our classifier on a validation corpus from February 2022, our analysis likely excludes many different types of potentially abusive content. Furthermore, given the low precision of the classifier, our analysis likely includes a number of false positives. This motivates future work that might identify a broader swath of content using extensive human labeling, surveys, qualitative interviews, and/or other methods.

A second notable limitation of our work is that it only analyzed impressions on Tweets with English-language text. A more complete analysis would require working with a wide range of languages and content types (pictures, emojis, videos, etc.) across all Tweets. This also means that our analysis does not capture some accounts' overall experience in their home timeline. Namely, some accounts in our study likely saw many Tweets from various different languages in their timelines, but our analysis only captures their exposure to English-language Tweets. It is also possible that some accounts in our data may be automated bots, though we do not control for that in this paper.

Our analysis of algorithmic impact uses an experiment at Twitter referred to as the "timelines quality holdback experiment" (Huszár et al. 2022). Some potential limitations apply to this experimental setup. First, it is applied at the account level rather than the session level (the experiment randomly allocates some accounts to only see reverse-chronological timelines, and some to see the standard algorithmically-ranked timelines). Furthermore, users in the algorithmically-ranked group can choose to see reverse-chronological timelines, although this is rare. Also, algorithmically-ranked timelines include multiple algorithmic services and machine learning models that influence the timeline, and our analysis does not isolate the impact of these algorithms on an individual basis.

Finally, while our analysis includes some group measurements based on average daily impressions, it does not include any analysis of demographic-based groups based on geography, gender, ethnicity, or age, for example. Demographic group analysis could shed more light as to the concentration of marginally abusive impressions, as we highlight in the future work section.

Implications of Concentration

A key finding of this paper is that exposure to marginally abusive content is extremely concentrated. In short, while a small portion of accounts bore the brunt of marginal impressions, most saw very few of these Tweets. This finding reinforces some key points from prior work: when it comes to potentially harmful content online, user experience varies widely. As has been the case for other types of harmful content (Grinberg et al. 2019), rarity at the population-level often obscures acute impact for some groups of users.

The impact of the observed concentration is difficult to reason about without additional context. In many settings, such as measurements of income or engagement, an equitable distribution is desirable. However, in the case of exposure to marginally abusive content, it is impossible to say whether higher or lower concentration is a better outcome without understanding the level of harm that a particular piece of content might cause for a reader. Reducing the overall volume and rate of exposure will necessarily make the distribution more concentrated; in the extreme case, having only one marginal impression across all users would be the most concentrated distribution possible. We thus suggest that future work on marginally abusive content not only focus on the volume and concentration of exposure, but also take on a perspective of harm reduction and prioritize users who are most impacted by marginally abusive impressions.

Even without more precise information regarding which groups are impacted by marginally abusive impressions, the observed concentration raises some potential concerns. For example, accounts in the 99th percentile (in terms of marginal impression rate) averaged 1 marginally abusive impression for every 16 total impressions on English Tweets. At Twitter's scale, this includes millions of people who experience different communication norms on the platform, and research suggests that these people may be more likely to create marginally abusive content (Brady et al. 2021). They also may be more likely to avoid the platform altogether.

Nuances in Algorithmic Impact

Our analysis illustrates how the impact of algorithmic systems can be multifaceted. By comparing marginally abusive impressions for accounts with reverse-chronological timelines, we found that accounts with algorithmically-ranked timelines experienced lower marginal impression rates, but higher cumulative volumes of marginally abusive impressions.

This nuanced finding offers a concrete example of the complexity in evaluating algorithmic systems. Some recent regulatory efforts characterize algorithmic feeds as generally harmful, and suggest that chronological feeds could reduce the associated harms (Gold 2021). However, our analysis shows the situation is more complicated; namely, using a reverse chronological timeline does not definitively improve a user's experience with marginally abusive content. Rather, our study adds to a growing body of literature that depicts a complex media ecosystem with an interconnected web of influential factors: content supply and demand (Munger and Phillips 2022), platform design/architecture (Malik and Pfeffer 2016), social network structure (Gallagher et al. 2021), and more, in addition to any effects from algorithms (Bandy and Diakopoulos 2021).

As Donella Meadows puts it in *Thinking in Systems*, "a systems insight... can raise more questions" (Meadows 2008). In this case, our results open up further questions about algorithmic timelines and exposure to marginal content: which groups see the highest rates of marginally abusive content? Which groups see the lowest rates? When do users tend to encounter marginally abusive Tweets? What other factors impact exposure patterns? Which factors should be addressed by the platform, and which might be addressed by users? While our work provides initial descriptive results and demonstrates the effect of algorithmically-ranked timelines, we hope these other questions can be explored in future research.

Future Work

Future related work may build on our analysis in several ways. Broadly speaking, other methods could be used to identify marginally abusive content, perhaps leveraging human annotation or improved models for classification. More granular analyses might also explore specific types of marginally abusive content (e.g. insults toward particular groups), and/or different types of margin content altogether (e.g. links to junk news and other low-quality content).

The variation in marginally abusive impressions also calls for future work in several directions. While our study analyzes impressions from across Twitter's platform, future work might analyze how specific surfaces (e.g. trends, moments, searches, profile visits, etc.) and Tweet types (e.g. retweets, replies, etc.) contribute to marginal impressions. We also emphasize that more work is needed that specifically focuses on the most-impacted users, asking which demographic groups experience the most marginally abusive impressions (e.g. based on age, gender, region, etc.), when marginally abusive impressions happen, and who creates marginally abusive Tweets in the first place.

Our analysis of algorithmically-ranked timelines also hints at some promising areas for future work in analyzing algorithmic systems. Given the multifaceted impacts we identified, research could explore users' perception of algorithmically-ranked timelines and the way they increase overall activity. Platforms might also explore methods to reduce marginally abusive impressions for more active users.

Finally, our results suggest future research efforts should avoid analyzing algorithmic impact as monolithic. Rather, the effects of algorithmic systems will be just as diversified and complex as the people using them.

Conclusion

In conclusion, our analyses contribute some of the first large-scale, empirical results for understanding exposure to marginal content on Twitter. We use real-world impression data on English Tweets from the full month of November 2021. Descriptive results show that exposure to marginally abusive content varies significantly across Twitter accounts, with a small group of accounts bearing the brunt of marginal impressions. Furthermore, experimental evidence shows that algorithmically-ranked timelines reduce marginal impression rates, but result in higher cumulative volumes of marginally abusive impressions due to increased activity.

Looking ahead, the paper points to several promising topics and directions for research. Future work might focus on demographic groups, for example, or isolate various algorithmic systems, actors, networks, and other factors that influence exposure to marginally abusive content. These research efforts will be critical to help understand and improve public communication on digital platforms.

Acknowledgments

We are grateful for colleagues who generously lent their time and expertise to help strengthen this project. In particular, we acknowledge Amanda Bower, Kyra Yee, Rumman Chowdhury, Luca Belli, Kristian Lum, Paul Lee, Kathy Yang, Lauren Fratamico, and the anonymous reviewers who offered feedback in the ICWSM review process. Without their work and support, this project would not have been possible.

Ethics Statement

This study intersects with a number of topics related to the ethics of algorithmic platforms. For example, our analysis requires collecting data about which Tweets users view while using Twitter, and also requires human annotators to review and rate potentially abusive content. Overall, we agree with researchers who view this type of work as necessary for understanding and protecting democratic discourse (Fiske 2022), especially in terms of standard risk-benefit frameworks. Still, it is important to note different measures taken to address potential risks.

While the holdback experiment is necessary, it is not ideal for many users to be excluded from algorithmic timeline features. Twitter has thus worked to provide more users access to algorithmic timelines while maintaining statistical robustness in the holdback experiment. As of 2020, the experiment included over 2 million active accounts in the reverse-chronological timeline group (Huszár et al. 2022), but when this analysis was conducted in 2021, that number had been reduced to 630k.

This study was not subject to an academic IRB process, however, it went through standard legal and privacy review processes at Twitter. Finally, the data used in this paper was fully anonymized for publication, following standard ethical procedures. We do not include any results that might disclose the identity of any account in the datasets.

References

Alexander, J. 2019. YouTube claims its crackdown on borderline content is actually working. *The Verge*.

Amnesty International. 2018. Toxic Twitter: A Toxic Place for Women. https://www.amnesty.org/en/latest/research/

2018/03/online-violence-against-women-chapter-1-1/. Accessed: 2023-04-19.

Auxier, B. 2020. 64% of Americans say social media have a mostly negative effect on the way things are going in the US today. *Pew Research Center.*

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239).

Bandy, J.; and Diakopoulos, N. 2021. Curating Quality? How Twitter's Timeline Algorithm Treats Different Types of News. *Social Media+ Society*, 7(3).

Bechmann, A.; and Nielbo, K. L. 2018. Are we exposed to the same "news" in the news feed? An empirical analysis of filter bubbles as information similarity for Danish Facebook users. *Digital journalism*, 6(8).

Bellmore, A.; Calvin, A. J.; Xu, J.-M.; and Zhu, X. 2015. The five w's of "bullying" on twitter: who, what, why, where, and when. *Computers in human behavior*, 44.

Brady, W. J.; McLoughlin, K.; Doan, T. N.; and Crockett, M. J. 2021. How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33).

Bruns, A. 2019. *Are filter bubbles real?* John Wiley & Sons.

Butler, A. P.; and Parella, A. 2021. Tweeting with consideration. https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration. Accessed: 2023-04-19.

Caplan, R.; and Gillespie, T. 2020. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society*, 6(2).

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3175–3187.

Constine, J. 2018. Facebook will change algorithm to demote "borderline content" that almost violates policies. *TechCrunch*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Farris, F. A. 2010. The Gini index and measures of inequality. *The American Mathematical Monthly*, 117(10): 851–864.

Feuston, J. L.; Taylor, A. S.; and Piper, A. M. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1).

Fiske, S. T. 2022. Twitter manipulates your feed: Ethical considerations. *Proceedings of the National Academy of Sciences*, 119(1).

Gallagher, R. J.; Doroshenko, L.; Shugars, S.; Lazer, D.; and Foucault Welles, B. 2021. Sustained online amplification of COVID-19 elites in the United States. *Social Media+ Society*, 7(2).

Gillespie, T. 2018. *Custodians of the Internet*. Yale University Press.

Gini, C. 1912. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E)*.

Gold, A. 2021. Exclusive: New bipartisan bill takes aim at algorithms. *Axios*.

Gray, M. L.; and Suri, S. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425).

Huszár, F.; Ktena, S. I.; O'Brien, C.; Belli, L.; Schlaikjer, A.; and Hardt, M. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1).

Katsaros, M.; Yang, K.; and Fratamico, L. 2021. Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Kwak, H.; Blackburn, J.; and Han, S. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI)*.

Lorenz, M. O. 1905. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70): 209–219.

Malik, M. M.; and Pfeffer, J. 2016. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media*.

Matias, J.; Johnson, A.; Boesel, W. E.; Keegan, B.; Friedman, J.; and DeTar, C. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018*.

Meadows, D. 2008. *Thinking in Systems: A Primer*. Chelsea Green Publishing.

Mozilla Foundation. 2021. YouTube Regrets: A crowdsourced investigation into YouTube's recommendation algorithm. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf. Accessed: 2023-04-19.

Munger, K.; and Phillips, J. 2022. Right-wing YouTube: a supply and demand perspective. *The International Journal of Press/Politics*, 27(1).

Nagar, Y. 2012. What do you think? The structuring of an online community as a collective-sensemaking process. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*.

Nurik, C. 2019. "Men are scum": Self-regulation, hate speech, and gender-based censorship on Facebook. *International Journal of Communication*, 13: 21.

Orellana-Rodriguez, C.; and Keane, M. T. 2018. Attention to news and its dissemination on Twitter: A survey. *Computer Science Review*, 29.

Redmiles, E. M.; Bodford, J.; and Blackwell, L. 2019. “I just want to feel safe”: A Diary Study of Safety Perceptions on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Roberts, S. T. 2019. *Behind the screen*. Yale University Press.

Scheuerman, M. K.; Jiang, J. A.; Fiesler, C.; and Brubaker, J. R. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2).

Starbird, K.; Arif, A.; and Wilson, T. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

The Economist. 2021. According to Twitter, Twitter’s algorithm favours conservatives. https://www.economist.com/graphic-detail/2021/11/13/according-to-twitter-twitters-algorithm-favours-conservatives. Accessed: 2023-04-19.

Vitak, J.; Chadha, K.; Steiner, L.; and Ashktorab, Z. 2017. Identifying women’s experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

Vogels, E. A. 2021. The state of online harassment. *Pew Research Center*, 13.