Measurement and Metrics for Content Moderation: The Multi-Dimensional Dynamics of Engagement and Content Removal on Facebook

Laura Edelson,* Borys Kovba,* Hanna Yershova, Austin Botelho, Damon McCoy, and Tobias Lauinger

Abstract. All major social media platforms use content moderation as a tool to prevent harmful content from spreading on their systems. To quantify the impact of content moderation, we propose the metric of *prevented dissemination*. To understand the practical limitations that content moderation systems face, we conducted an empirical measurement study of public posts from news providers on Facebook in English, Ukrainian, and Russian. We analyzed how quickly posts accrue engagement, finding large asymmetries of engagement over content and time, and use our measurements to build a model to predict a post's future engagement. We also observed the timing of (rare) post removals. Using our prevented dissemination metric, we estimate that removals prevented only 24–30% of the posts' predicted engagement. Our lens of prevented dissemination provides an outcome-based metric to judge the impact of content moderation in practice and could help builders of moderation systems prioritize content for review.

1 Introduction

Content moderation has become an unavoidable requirement for any major social media platform (Gillespie 2018) to prevent it from being inundated with illegal, hateful, or otherwise undesirable content. Content moderation has multiple facets, including both hard moderation of removing posts that violate stated content policies (Meta 2024b), and "soft moderation" of decreasing the reach of "borderline" content that nearly (but not quite) breaks a platform's rules (Zuckerberg 2018). Moderation can be controversial, not only because of what is written in platforms' content policies but also because of how these policies are enforced. Platforms are frequently criticized for moderating both

*. These authors contributed equally to this work.

doi:10.54501/jots.v2i5.220

too little (Color Of Change 2020) and too much (Taibbi 2022). To show that they are "making progress," platforms now publish transparency reports with some details about their content moderation efforts (Meta 2024a). However, these transparency reports have themselves come under frequent criticism for being "confusing, uninformative, and evasive" (Sorensen 2021), and civil society organizations regularly call on social media companies to make more detailed information about their moderation systems publicly transparent (Integrity Institute 2024; MacCarthy 2020).

A specific point of criticism is that while these reports disclose how much content was moderated (e.g., number of posts removed), it is still unclear what *difference* the content moderation efforts make in practice (e.g., how many people were not exposed to violative content) due to the limitations of the metrics being reported. This lack of clarity is because, as prior work has shown, only a small number of posts will become widely seen, and posts have relatively short life cycles (Goel et al. 2015; Alhabash and McAlister 2015; Shin et al. 2018). This means that which posts are moderated and when they are moderated can have large effects on user exposure.

What becomes widely seen (and when) depends not only on "natural phenomena" such as audience preferences, but also on platform implementation choices, such as which content the recommendation algorithm shows to users (and when). The dimensions of popularity across content and the evolution of popularity over time have been studied before in isolation. However, their interaction and implications on the effectiveness of content moderation have not been explored well. While it is intuitive that content moderation should be fast, because removing content after interest has already subsided will have very limited benefit, we are not aware of any metric that operationalizes this intuition or a study that quantifies the "cost" of slow moderation. Judging whether platforms are increasing the efficacy of their content moderation requires an appropriate metric, and it appears unlikely that platforms would propose such a metric voluntarily if it risks shining a bad light on their existing efforts.

To provide context for such a metric and explore the practical realities that content moderation systems must contend with, we first study the asymmetry of user engagement(such as commenting, resharing, or reacting) over content and over time in Section 5. We find that in each of the three language-based content ecosystems we study, there are wide variations of engagement across posts, as well as short life cycles of user engagement. For example, in our dataset of over 2.6 M public news posts in US English, Ukrainian, and Russian, the top 1% most-engaged content was responsible for 58% of user engagement in US English, 45% in Ukrainian, and 57% in Russian news media Facebook content. Across all posts in the three languages, the median time to 50% engagement is just 3 hours.

These findings underscore that in terms of user exposure to content, it matters both *which* posts are moderated (whether they have the potential to reach a large audience), and *when* they are moderated (whether they have already reached their potential audience).

To capture both of these realities, we propose the metric *prevented dissemination* as a measure of a moderation system's effect on the user experience. Our metric, as defined in Section 6, does this by quantifying the difference between an estimated (or observed) total dissemination of content without intervention and the reduced dissemination with a soft or hard moderation intervention.

To demonstrate the utility of this metric, in Section 7, we employ it in a case study to measure the impact of content removal events that occurred during our study period. In order to infer how quickly posts are removed in practice, we periodically probe whether the posts in our three datasets still exist (between 4 and 12 times per day, depending on the dataset). Our data source does not directly indicate which content was removed, or the reason for removal. Therefore, we have no certainty whether observed removals were instances of content moderation or voluntary removals. Either way, the case study illustrates how the prevented engagement metric can be used to quantify the impact of removals and to compare hypothetical scenarios (e.g., sooner vs. later removals). To estimate a post's potential future audience (and calculate the prevented dissemination metric), we build a model to predict a post's future engagement and apply it to the removed posts. We estimate that due to their timing, the observed removals prevented only 24.3% of the engagement that the US English posts would likely have received within 48 hours had they not been removed (30.5% in Ukrainian). Removals of Russian posts occurred late enough in their respective life cycles that, on average, no engagement was prevented. Overall, the metric and our analysis of the metric in practice provide a useful way to assess the potential impact of content moderation and could help system builders improve their content moderation pipelines.

2 Background

Every social media company that we are aware of publishes written standards for what kinds of content users are and are not allowed to post and what content is or is not eligible to be recommended in algorithmic feeds e.g. (Meta 2024b; YouTube 2024; TikTok 2024b). There is a great deal of variability in these content policies, which have been the focus of much scholarly work (Buckley and Schafer 2022; Gillespie 2018).

The current generation of social media applications are nearly all driven by algorithmic feeds (Meta 2024f; TikTok 2020; Twitter 2023; Goodrow 2021). Even older platforms, such as Facebook and Reddit, have added algorithmic feeds to their existing products (Meta 2024g; Reddit 2024). Many social media platforms, including Facebook, have moved toward employing a more content-based feed algorithm that can distribute individual posts more widely and rapidly than social feeds (i.e., collaboratively filtered) or search-focused algorithmic feeds (Heath 2022). Similarly, when users primarily consume content in a feed, rather than manually navigating to different pages on a site or searching for content, a new piece of content can be shown to a very large audience

immediately.

The moderation systems that enforce content policies consist of a variety of multimodal review pipelines (Gorwa et al. 2020) that may review different subsets of posts for different reasons (such as user reports, keywords, or other signals) and may result in a variety of enforcement actions, some of which are externally visible (hard modaration such as content removal or user bans), and others which are not (soft moderation such as downranking or "shadowbanning") (Haugen, n.d.). Some review systems are entirely automated (Meta 2024d), some are entirely manual, and many combine automated and manual review components (Meta 2024c). Meta takes virality into account when prioritizing content for review but gives little detail about how they do this (King and Gotimer 2020). Meta also publishes "Transparency Reports" with data about absolute counts of removed posts in various categories of violation, as well as the proportion of overall views for selected categories of violation (Meta 2024a). A significant amount of research has also studied enforcement mechanisms at the end of these moderation review systems for a wide range of content categories, including child sexual abuse material (CSAM) (Farid 2018), fraud (Bian et al. 2021), and violence (Sudhakaran and Lanz 2017).

A particularly understudied area is the temporal dynamics of review systems and the enforcement actions that result from them, as well as the broader challenges to these systems that stem from the rapid and uneven dissemination dynamics of social media networks. It is anecdotally well known that social media content has a short shelf life, which means that *when* content removal happens (not only *what* content gets removed) has a significant impact on users' experience. However, we currently lack an operationalizable metric to evaluate the potential benefits of quicker review pipelines.

2.1 Transparency Reporting and Metrics for Content Moderation Systems

Many major social media platforms publish quarterly or semiannual transparency reports with information about their content moderation enforcement (Meta 2024a; TikTok 2024a; Google 2024; X 2024). All of these report the prevalence of violative content and what share of actions resulted from proactive review (as opposed to user reporting), broken out by category of violation. YouTube and TikTok also publish information about how many views posts received before their removal, although this number is hard to contextualize, absent an understanding of the overall distribution of views over content on a platform. TikTok additionally publishes limited information about when removals occur, publishing a "removed at 24 hours" statistic for different categories of violations.

Meta also publishes a quarterly transparency report with statistics about removals for various categories of content (Meta 2024a). This report says, "We remove millions of violating posts and accounts every day on Facebook and Instagram. Most of this happens automatically, with technology working behind the scenes to remove violating content—

often before anyone sees it. Other times, our technology will detect potentially violating content but send it to review teams to check and take action on it" (Meta 2024d). Meta describes three factors it uses to prioritize content for review: "severity, virality, and likelihood of violation" (Meta 2024c).

Multiple civil society organizations, including the Integrity Institute (a trade organization for trust and safety professionals), have called on all platforms, and Facebook specifically, to publish metrics that capture dissemination of policy-violating content (Integrity Institute 2024; Sorensen 2021). Specifically, the Integrity Institute has called for platforms to publish the "reach" of harmful content over several different periods, the time delay between posting and content removal, and the factors that most contributed to user exposure to violative content (Integrity Institute 2024).

In the absence of platforms reporting such metrics themselves, it may be possible to infer them from other data sources. For example, the European Union's Digital Services Act (European Commission 2023) regulates content moderation practices and requires very large online platforms to report metadata about moderated content to a public database maintained by the European Commission (European Commission 2024). While this database is a promising step forward by providing "ground truth" about content moderation and has already attracted scholarly interest (Drolsbach and Pröllochs 2024; Kaushal et al. 2024; Trujillo et al. 2024), it unfortunately lacks engagement data, making it unsuitable for our specific research question. Instead, in this work, we show how a transparency tool built for a different purpose (CrowdTangle) can be repurposed to estimate prevented engagement metrics externally. Unfortunately, Meta retired CrowdTangle in 2024 (Meta 2024e). The company has instead directed researchers to use its Content API (Clegg 2023), although we note that the data we use for our analyses in this work are not available through this tool.

3 Related Work

3.1 Empirical Measurement of Social Media

Nearly as long as social media platforms have existed, researchers have studied how users interact with content on these platforms. In recent years, more work has been dedicated to empirical measurements of the large differences in engagement between widely seen content and less popular content, although these analyses tend to be retrospective rather than observing engagement as it is happening. Vosoughi et al. (2018) study "rumor cascades" on Twitter and identify significant differences between true and false information, but do not study the ecosystem of content as a whole. In work related to ours (and from which we source a list of US news Facebook pages), Edelson et al. (2021) study distributions of user engagement with US news publishers and observe significant differences based on a partisanship and factualness, but they do not study engagement

over time. Mohammadinodooshan and Carlsson (2023) also find differences between these categories in how posts accrue user engagement over time. Lazovich et al. (2022) study engagement over all users for a large set of Twitter users and compare a variety of metrics (including top 1% and Gini coefficients) for explaining such heavily skewed distributions, while Pfeffer et al. (2023) explore how quickly tweets get impressions, finding that the mean time to 50% of first-day impressions is slightly greater than 2 hours. Most recently, McGrady et al. (2023) collect a large random sample of YouTube posts, observing a similarly highly skewed distribution of views and engagements on that platform. In a study of content from the top 1,000 Instagram accounts, Thorgren et al. (2024) observe that, on average, posts get 75% of their total engagement (likes and comments) in 13 and 7.5 hours after posting, respectively.

3.2 Predicting Engagement

In order to understand the impact of content removal on user engagement, we will first need to predict possible future engagement. Heiss et al. (2018) study the association between characteristics of political pages on Facebook and user engagement, finding correlations between engagement and page followers and post type (as we do) as well as real-world characteristics of posting accounts. Early work studying patterns of engagement with social media content studied "cascades" of user sharing and resharing of content, finding that early resharing activity can be highly predictive of later activity (Cheng et al. 2014). Because of the perceived importance of the phenomenon of virality in determining what content becomes widely seen, several approaches to predicting temporal patterns of engagement (Vassio et al. 2022), viral resharing (Jenders et al. 2013), and final engagement (Mohammadinodooshan and Carlsson 2023) have been advanced. These works also find that early patterns of engagement with content can predict later content engagement, and our findings echo this.

3.3 Study of Removed Content

In addition to content creation and engagement, researchers have qualitatively studied the impact of content moderation. Singhal et al. (2023) conducted a landscape analysis of content moderation practices. A study of suspicious Twitter account creations and suspensions by Pierri, Luceri, Chen, et al. (2023) also provides insights into the content of tweets removed as a result of account suspensions. Only indirectly related to content moderation because of a presumption that users were the source of content removals, several researchers have attempted to identify significant differences between deleted and non-deleted tweets as well as users who did and did not have removed tweets (Zhou et al. 2016; Bhattacharya and Ganguly 2016; Almuhimedi et al. 2013-02-23). These studies found that deleted posts were more likely to contain "regrettable" content and also found significant differences between users who did and did not have removed content. More recent work studying content moderation of COVID-19-related misinformation by Papakyriakopoulos et al. (2020) finds that once posts with links to known

misinformation are removed, other posts linking to the same webpages decrease significantly. Pierri, Luceri, Jindal, et al. (2023) study Russian propaganda and low-quality content about the conflict during the early period of the invasion of Ukraine in Russian, US English, and Ukrainian. Their work also studies removed content by category but does not attempt to measure when content is removed. Another line of research aims to identify and characterize the users behind low-credibility content. For example, DeVerna et al. (2024) use structural graph features to identify "superspreaders" on Twitter and assess the (hypothetical) impact on the availability of low-credibility content if these users are removed. The motivating assumption behind this approach is that prioritizing for intervention the most prolific sources of policy-violating content could have a large platform-wide impact. Our research question is orthogonal by focusing on the temporal aspect and treating all content independently: It yields insights into practical constraints for everyday content moderation as opposed to targeted intervention against repeat offenders. To our knowledge, we are the first to empirically measure *when* content removal happens and estimate the impact of those removals on user engagement.

4 Data Collection and Datasets

In order to provide concrete data about constraints that patterns of dissemination have on content moderation review systems, we collected empirical data about user engagement with news content on Facebook. Our methodology involved first identifying news providers on Facebook so that their public content can be retrieved, then frequently measuring how much engagement it receives over time, and finally detecting if and when such content is removed.

4.1 Identifying US English, Ukrainian, and Russian News and Media Publishers on Facebook

For our data collection, we focused on news providers in a wide sense because their content is public, can have high user exposure, may be subject to controversy (and therefore be moderated), and can be delineated from other types of content (to balance completeness of the studied ecosystem with a manageable dataset size). Unfortunately, there is no direct functionality to exhaustively search for news providers on Facebook given our criteria. We therefore based our measurement on a list of US news sources that we obtained from the authors of a study on user engagement with misinformation (Edelson et al. 2021), which was originally built using data obtained from NewsGuard (NG) (NewsGuard 2021) and Media Bias/Fact Check (MB/FC) (Media Bias/Fact Check 2021). In contrast to the prior study, we did not remove low-activity pages since our goal is to study the entire spectrum of content.

To ensure generalizability and avoid "overfitting" on aspects that may be unique to the US news ecosystem, we also included news providers and entertainment pages publishing

in a different geographical region and in a different language. We chose news in Ukrainian and Russian because of the ongoing war, the potential for propaganda or misinformation campaigns, and the corresponding need for content review and moderation. To our knowledge, there is no preexisting list of news sources published in Ukrainian or Russian languages comparable to the one we were able to obtain for the US context. Therefore, two Ukrainian- and Russian-speaking students collaborating with our research group comprehensively searched the CrowdTangle web interface for Ukraine war and politics-related terms (e.g., *war*, *soldiers*, *President Zelenskyy*) in Ukrainian and Russian in November 2022. At the time, the war was the dominant news topic in these languages, and the resulting list of Facebook pages contains a mix of traditional news providers and some entertainment-focused pages that had posted at least once about the war (or had mentioned celebrities linked to President Putin, for example) in either Ukrainian or Russian. Even though Facebook had already revoked service in Russia, many pages remained that served Russian-speaking audiences in Eastern Europe and elsewhere. In total, we used 10,469 US-centric news pages in US English and identified 4,675 Ukrainian-language and 2,909 Russian-language news (and entertainment) pages.

4.2 Collecting Data from CrowdTangle

We extracted data on these news pages' public posts (including posts not matching any of the initial page discovery keywords) using Meta's now-defunct CrowdTangle API (Shiffman 2020). Meta shut down CrowdTangle in 2024 (Meta 2024e), but prior to this, CrowdTangle gave researchers access to publicly viewable content on Facebook or Instagram. Data collection ran from July 1, 2023, to August 1, 2023, for the US dataset and from June 17, 2023, to July 17, 2023, for Ukrainian and Russian datasets.

Available data included the post content and metadata such as the post language, information about the corresponding Facebook page (such as the verification status (Meta 2023) and subscriber count), and a time series showing accrual of user engagement over time. User engagement data included counts of comments, shares, likes, and other types of reactions (e.g., "love" or "sad"). The API did not return deleted posts, thus we retrieved data periodically so that we could observe posts before they were deleted. We maintained separate data pipelines for each language grouping, including separate CrowdTangle dashboards, data collection processes, and databases. Given the number of pages that we monitored and the rate limits of the API, we are able to collect all posts published in the previous 7 days every 6 hours for our US English news pages and every 2 hours for Ukrainian and Russian media pages. We were able to poll Ukrainian and Russian media pages more frequently because there were fewer of these sources, meaning that each crawl took comparatively less time to run compared to a crawl of US English pages.

For posts that were not removed, we last observed their "total" engagement accrual as it was reported by CrowdTangle after 7 days. For posts that *were* removed, we can narrow down the removal time to a window of 6 hours for our US-centric English news pages, or

2 hours for Ukrainian and Russian news pages.

4.3 Identifying and Filtering by Post Language

In our analysis, we separate the three news ecosystems by the language of the posts. Our initial sourcing of pages on CrowdTangle for Ukrainian- and Russian-language media was intentionally broad and thus included pages that had posted Russian or Ukrainian war-related content on a handful of occasions but did not regularly post in those languages. Therefore, we additionally needed to filter posts from our raw collection for language relevance. While CrowdTangle included a language attribute for most posts, it was sometimes missing, especially for photo-type posts with no text. To minimize the impact of this data artifact on our analysis, we determined whether each page only posted in a single (known) language and propagated it to any post where the language attribute may be missing. The strong majority of pages we monitored were monolingual; only 3% had language-tagged posts in different languages. Virtually all of these pages featured a combination of Ukrainian- and Russian-language posts with varying proportions; only 4 pages showcased posts in US English and Russian. We excluded 124,849 posts because we could not determine their language and, therefore, could not assign them to a primary language dataset. After data cleaning, our dataset contains 2,687,314 posts from 17,504 unique pages.

4.4 Harmonizing Time Series Data

In our analysis, we seek to characterize how quickly posts accrue engagement over time, and for how long posts accrue engagement. We do so based on CrowdTangle's engagement time series data from the last observation of a post, which reports engagement over time in intervals that expand from initially 15 minutes after post publication to 6-hour time steps later on (Shiffman and Fan, n.d.). These expanding time series appear to be roughly scaled to normalize total user engagement between time steps.

Sometimes, CrowdTangle stopped reporting additional time steps 48 hours after posting (at this point the majority of posts cease to get any new engagement), and cut off updates for more posts as more time after post creation passed. Ergo, to ensure the consistency of measurement, we restrict our analysis to the first 48 hours of user engagement. The vast majority (83.5 %) of posts' "total" engagement (as reported by CrowdTangle after 7 days) is already accrued during this initial 48-hour period. We will refer to this as the "48H engagement," and consider it the 100% for the purposes of our analysis. We exclude two types of posts that do not have engagement time series data for the full 48 hours. First, 155 posts have entirely corrupted time series data, and ≈11% of posts, 347,444 in total, do not have time steps through the 48 hour mark. We observe no obvious temporal or page-based bias in which posts ceased reporting time steps before or after two days, thus we do not believe this pattern or our exclusion of these posts to bias our analyses.

The time step counts, gaps, and timings are not identical across posts. We do not find any meaningful correlation between 7- or 2-day engagement and these time step features, meaning they do not bias our analysis. However, to ease comparisons between posts, we quantize the engagement time series to one value every 30 minutes. We tested linear interpolation between the two nearest time steps and a curve-fitting approach. Both yielded comparable accuracy, thus we opt for the simpler linear interpolation.

4.5 Identifying Removed Posts

During our data collection, we observed that 16,158 posts were removed before the end of the 48-hour period. Removals are not explicitly marked in the API, thus our visibility is very limited. We can only *infer* that a post was removed when it was observed in one iteration of our data collection but not in any later iteration. We confidence tested the fact that posts that disappeared from CrowdTangle reflected real removals of posts from Facebook by manually validating a random sample of 50 removed posts and confirming that they were, in fact, no longer accessible. Our data collection cycles repeated 4 times per day for our US collection, and 12 times per day for our Ukrainian and Russian collection. We do not know *precisely* when removals occur, but we can narrow their removal to the 6-hour window between collections for our US-centric dataset, and a 2-hour window for our Ukrainian and Russian datasets. To err on the side of caution, we estimate the lifetime of removed posts as the time between publication and the last time the post was observed in our data collection; that is, we make the conservative assumption that a post was removed immediately after we last observed it. (In reality, it could have remained available up to 6 hours longer in our US dataset or 2 hours longer in our Ukrainian and Russian datasets.) In our three datasets, we detect 12,864 deleted posts in US English, 1,071 deleted posts in Ukrainian, and 2,223 in Russian. We do not analyze the content of deleted posts, but only study their metadata, notably how long these posts remained accessible, and how much engagement they accrued.

Our analysis of deleted posts is subject to two important caveats. First, we do not know *who* removed a post and cannot distinguish authors taking down their own post or Facebook moderating the content. Second, we only detect deleted posts that appeared in our dataset at least once. Most importantly, this implies that our analysis does not include posts removed before publication (e.g., moderated by an automated detection system at time of posting). We also likely undersample posts removed very quickly after publication because of the intermittent nature of our crawl. To quantify this potential undersampling, we compute a post's "collection lag" as the difference between its publication timestamp and the time it first appears in our crawl. The lag is not constant across posts due to varying delays until new posts become visible in the API and variability in the timing of our data collection. We compare the collection lag distributions of removed and non-removed posts, and find only small effects: removed posts had on average a 9% shorter lag in the US English dataset, and a 7.5% shorter lag for the Ukrainian and Russian collection. Under ideal circumstances, the lag distributions of removed and non-removed posts

Table 1: Summary statistics for the post language datasets after data cleaning.[2]

Primary Dataset	US English	Ukrainian	Russian
Pages	10,447	4,673	2,893
Posts	1,762,535	500,934	423,845
Time Steps	49,431,257	14,049,924	11,877,889
Removals Dataset	**US English**	**Ukrainian**	**Russian**
Posts	12,864	1,071	2,223

would be similar; the difference arises from removed posts with a longer lag and a lifetime shorter than the lag, which prevented them from being observed and included in our dataset. Overall, this modest difference shows that the dataset of removed posts is only slightly (but not heavily) biased towards longer-lived posts.

4.6 Datasets for Analysis

In the remainder of this work, we will analyze two datasets, our *Primary* dataset of posts for which we have observed a full 48 hours of data, and our *Removals* dataset, made up of posts that appear to have been removed during our observation window. We will primarily (but not exclusively) analyze these two datasets further broken down by language. Table 1 presents summary statistics for each of these two datasets.

5 Analysis of Dynamics of User Engagement With Content

In this section, we empirically study content dissemination in three disjoint language-based news and media ecosystems on Facebook in order to better understand what practical constraints they may create for content moderation systems. Details on how we sourced these content datasets are presented in Sec. 4.1. We first explore the asymmetry of engagement across posts, and establish the classes of engagement size that will serve as the primary lens for our further exploration of engagement over time. We then proceed to study the factors correlated with higher classes of engagement, and finally study the speed at which posts accrue engagement.

5.1 Distribution of Engagement over Posts

We begin by analyzing how engagement is distributed over posts. This informs how much review capacity is needed to cover a certain share of the audience. When engagement is very skewed toward a small number of posts, for example, the system may be able to achieve higher coverage by reviewing fewer posts than when engagement is

2. There is a slight overlap of pages because a page can post in more than one language (e.g., some posts in Ukrainian, some in Russian). The total number of unique pages is given in Section 4.3.

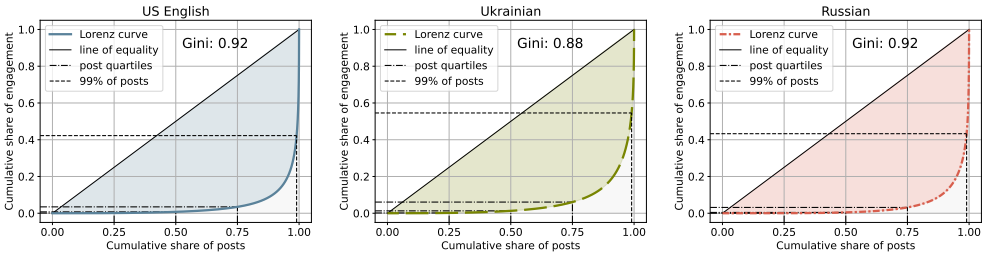

Figure 1: Lorenz curves of post engagement in the US English, Ukrainian, and Russian data sets. The black 45˚ lines represent hypothetical engagement equality and contrast how skewed the real distributions are. Most posts receive very little engagement, which is concentrated among the most engaging posts.

distributed more evenly. Common tools to measure inequality (Gastwirth 1972) (e.g., income inequality) in economics are Lorenz curves (Gastwirth 1971)[3] and the Gini coefficient (Farris 2010).[4]

As seen in Figure 1 by the deviation from the diagonal line (which represents no inequality), the posts in all three language datasets show extremely skewed distributions of engagement. The corresponding Gini coefficients are 0.92 for US English, 0.88 for Ukrainian, and 0.92 for Russian on a scale from 0 to 1, indicating extreme inequality in all three distributions. By another measure, the top 1% of posts ordered by engagement would cover 58% of user engagement in US English, 45% in Ukrainian, and 57% in Russian. Even higher inequality has been observed on other platforms. On X/Twitter, the top 1% of authors receive ≈80% of all views, and the Gini coefficients of views and a variety of engagement types always exceed 0.95 (Lazovich et al. 2022). On YouTube, in a random sample of videos, the top 0.16% accumulate 50.52% of all views (McGrady et al. 2023). This means that despite the apparently high degree of inequality of engagement over content, the news ecosystems we observe on Facebook have relatively "flatter" distributions of content than some other social media content ecosystems.

Post classes. In our analysis, we will explore how different parts of these distributions behave very differently over time. Therefore, treating the distributions of engagements over posts as single monoliths will obscure key insights. Instead, we separate posts into five (unequal) classes based on their 48H engagement so that differences between them are easier to ascertain. This also allows us to focus some analyses on the most impactful class in terms of engagement, which we denote by P99. This is the 99th percentile of posts by 48H engagement (i.e., the top 1%). The remaining four classes are based on the quartiles Q1–Q4, with two exceptions: In the Russian language dataset, 26% of posts received no engagement, so its Q1 actually contains 26% of posts. Furthermore, we exclude the top 1% from the Q4 classes so that the five classes are non-overlapping.

3. Lorenz curve: Cumulative share of total dataset engagement with posts that have been ordered by per-post engagement.

4. Gini coefficient: Average absolute difference between all items in a set divided by the overall set average value.

Table 2: Mean, median, and maximum 48H engagement for posts in the Q1–Q4 and P99 engagement classes. Maximum values for the Q1–Q4 classes also function as the cutoff between classes. The top 1% (P99 class) contains the bulk of engagement, and exhibits the largest difference between the three datasets.

Dataset	US English	Ukrainian	Russian
Q1 avg	1	1	0
median	1	1	0
maximum	3	3	0
Q2 avg	8	9	2
median	8	8	2
maximum	14	16	4
Q3 avg	33	37	12
median	29	33	11
maximum	66	71	27
Q4 avg	479	386	188
median	205	200	88
maximum	4,495	2,745	1,590
P99 avg	16,978	8,726	6,291
median	8,644	4,915	3,254
maximum	650,984	213,634	144,109

Average, median, and maximum engagements for posts in the five classes of engagement are shown in Table 2. In addition to having fewer posts overall, posts in the Ukrainian and Russian datasets also have fewer engagements on average. Overall, the Ukrainian and Russian datasets have 192 and 111 average engagements per post, compared to 294 on average for posts in our US English dataset. While the scale of the three datasets is not the same, the bulk of the difference in engagement between the three languages is at the top of the distribution in the P99 class. The remainder of the distributions, particularly the Q1–Q3 classes, are broadly similar in terms of engagement across datasets.

As noted in Section 4.4, engagement continues to accrue after the 48-hour period, and for a small percentage of posts, the increase is nonnegligible. However, despite the increase in engagement, the vast majority (97%) of posts did not change their class between the 2-day and 7-day mark. This class stability further motivates our use of classes in this analysis.

5.2 Factors Correlated with Engagement

For a content moderation review system, it would be useful to be able to predict how popular a post might become so that the system can use expected future engagement to prioritize among otherwise equivalent posts. As the first step toward predicting user engagement, as we will do in Section 7.1, we test several page and post features for correlation with 48H engagement. This is useful for understanding how predictable engagement is prior to a post's initial dissemination. In detail, we look for associations between a post's *class* of 48H engagement and how many subscribers a posting page has, whether the posting page was verified or not, whether or not the post was a video, the time of day of posting, and the day of week of posting. We use Kendall's τ test to

Table 3: Correlations between page or post features and the 48H engagement class of posts. Significant correlations are in bold.

Data Set	US English	Ukrainian	Russian
Subscriber Quartile	**Moderate (τ = 0.41)**	**Weak (τ = 0.22)**	**Moderate (τ = 0.31)**
Page Verified	**Weak (τ = 0.23)**	**Weak (τ = 0.24)**	**Weak (τ = 0.14)**
Video Media Type	**Weak (τ = 0.08)**	**Weak (τ = 0.10)**	**Weak (τ = 0.17)**
Time of Posting	no	no	no
Day of Posting	no	no	no

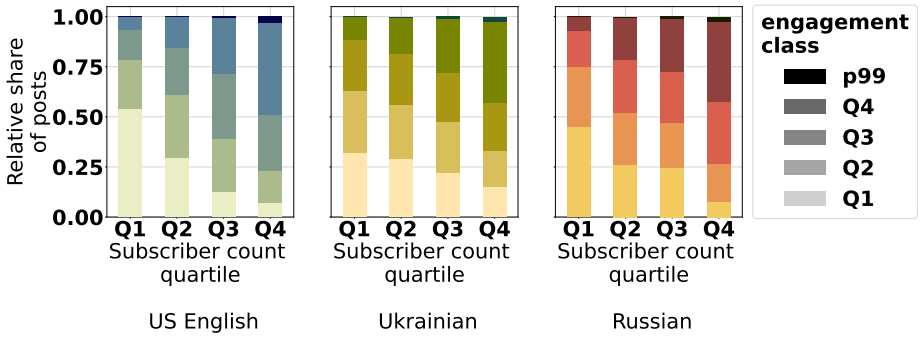

Figure 2: Distribution of engagement class by subscriber count quartile. Pages with more subscribers tend to have more posts in the high engagement classes.

measure the correlation between these different features and the post's engagement class, adjusting p-values with the Bonferroni correction (Armstrong 2014). When testing for correlation between subscriber count and 48H engagement, we additionally assign each page a subscriber count quartile value.

The strongest correlation we find is between the post's engagement class and the corresponding page's subscriber quartile, although the degree of correlation differs. For US English and Russian posts, Kendall's τ coefficients are 0.41 and 0.31, respectively, indicating a moderate (Wicklin 2023) correlation; for Ukrainian posts, it is 0.22, indicating a weak correlation. Figure 2 visualizes the distribution of post engagement class by page subscriber quartile and shows how higher subscriber count quartiles have more high-engagement class (Q4 and P99) posts. We performed additional testing for correlation between a post's absolute engagement number and absolute page subscribers and found similar results. Figure 3 presents a log-scaled box plot of absolute engagement and shows that the relationship between subscribers and engagement was consistently positive: higher subscriber count quartiles exhibited greater median and mean engagement across all language ecosystems. It is interesting to note that this is true even though, according to Meta's Widely Viewed Content report from the time period of our study, "Page Follows" accounted for no measurable share of recommended content in users' feeds, meaning this engagement is not directly attributable to these subscriptions (Meta 2024g).

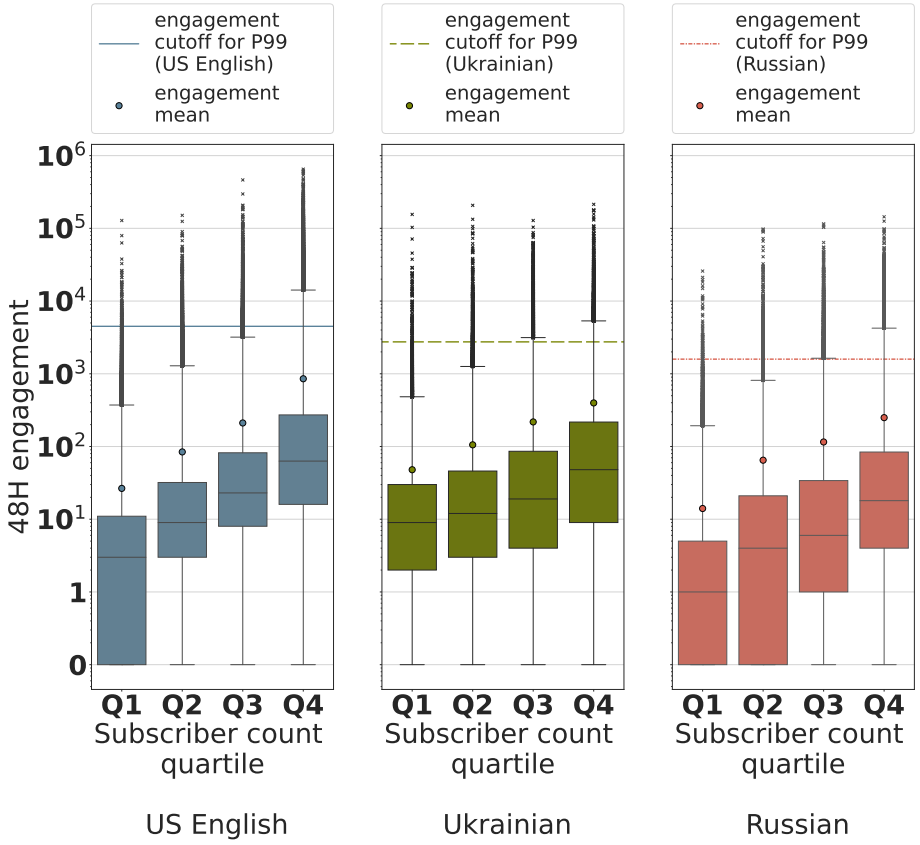

Figure 3: Box plot of distribution of engagement by subscriber count quartile (log scale). Whiskers enclose 0–99th percentiles. Pages with more subscribers tend to have higher median and mean engagement per post.

In terms of a page's verified status, we find a weak correlation with the post-engagement class for all three languages (τ = 0.23, τ = 0.24, and τ = 0.14). This is visualized in Figure 4. Figure 5 also shows a log-scaled box plot of absolute engagement by page verified status for each language, and shows the higher median and mean average engagement that verified pages have. Finally, we also find a weak correlation between a post's media type and post-engagement class (non-video/video) for all three languages as well, with video posts more likely to be higher engagement (τ = 0.08, τ = 0.10, and τ = 0.17).

Prior work studying the Facebook pages of Italian influencers has found a correlation between post engagement and the time of day and day of week of posting (on a per-page basis) (Vassio et al. 2022). In our datasets, using 48H engagement and comparing across all pages, we do not find any such correlation. That is, on a larger scale, the posting time does not appear to have a significant influence on engagement.

5.3 Speed of Post Engagement

To minimize engagement with policy-violating posts, reviewing them before most engagement accrues is critical. To that end, we analyze how quickly individual posts accrue

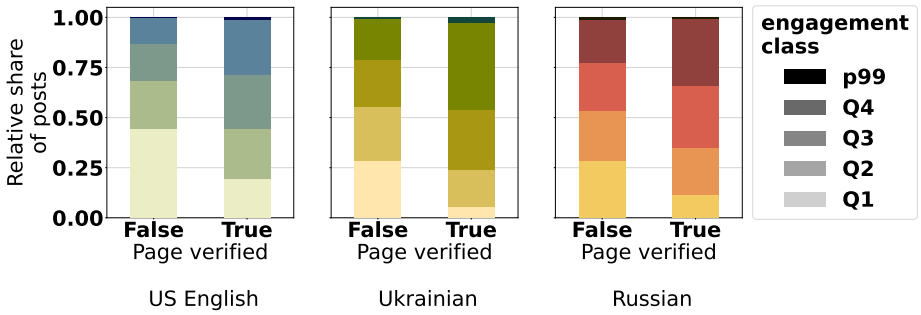

Figure 4: Distribution of engagement class by page verified status. Verified pages tend to have more posts in the high-engagement classes.

engagement over time. We measure the *velocity* of accrued engagement in terms of engagements per hour (EPH).

Figure 6 shows mean accrued engagement over the two-day observation period for the P99 engagement classes, and Figure 7 plots the corresponding mean engagement velocity in the three datasets. We do not show the other classes in these visualizations because they are indistinguishable from the x-axis when the y-axis is linearly scaled. While the scales of velocity differ between the three languages, the curves follow a generally similar trend. Velocity is typically highest immediately after posting, and within the first 2–5 hours, it quickly declines to a level that is multiple times smaller. For the P99 class, this intermediate plateau is still high in absolute terms (at a mean of just below 400 engagements per hour for US English posts), which remains relatively stable for 10–20 hours. In the P99 class, the plateau is also followed by a slight "second-day" effect, when velocity briefly increases approximately 24 hours after the time of post creation before again declining more gradually over the remaining duration of our 48-hour observation window.

As prior work on other platforms has done (Pfeffer et al. 2023), we find the half-life of post engagement, defined in our dataset as the time at which a post accrues 50% of its 48H engagement, to be a useful metric for numerically comparing the speed of engagement accrual. The scale of engagement accrual highlights that effective content moderation delays need to be measured in hours, not days. For example, if posts in the P99 class are to be reviewed before they have accrued half of their 48H engagement, on average, the review must be completed by hour 13.6 for US English posts, by hour 12 for Ukrainian, and hour 16.2 for Russian. We note that applying the same standard to the Q1–Q4 classes can be more challenging for a moderation review system, as these posts have an even shorter 48H average engagement half-life of 5.4 hours in US English, 4.9 hours in Ukrainian, and 5.1 hours in Russian. In absolute terms, US English posts in the P99 class take a median of only 1.5 hours (average: 4.5 hours) to reach 1,000 engagements or a median of 16.5 hours (average: 18 hours) to reach 10,000 engagements. (Not all P99 posts make it to 10,000 engagements.) While there are few posts of this level of activity (1%), their speed presents a challenge to review the content before a substantial amount

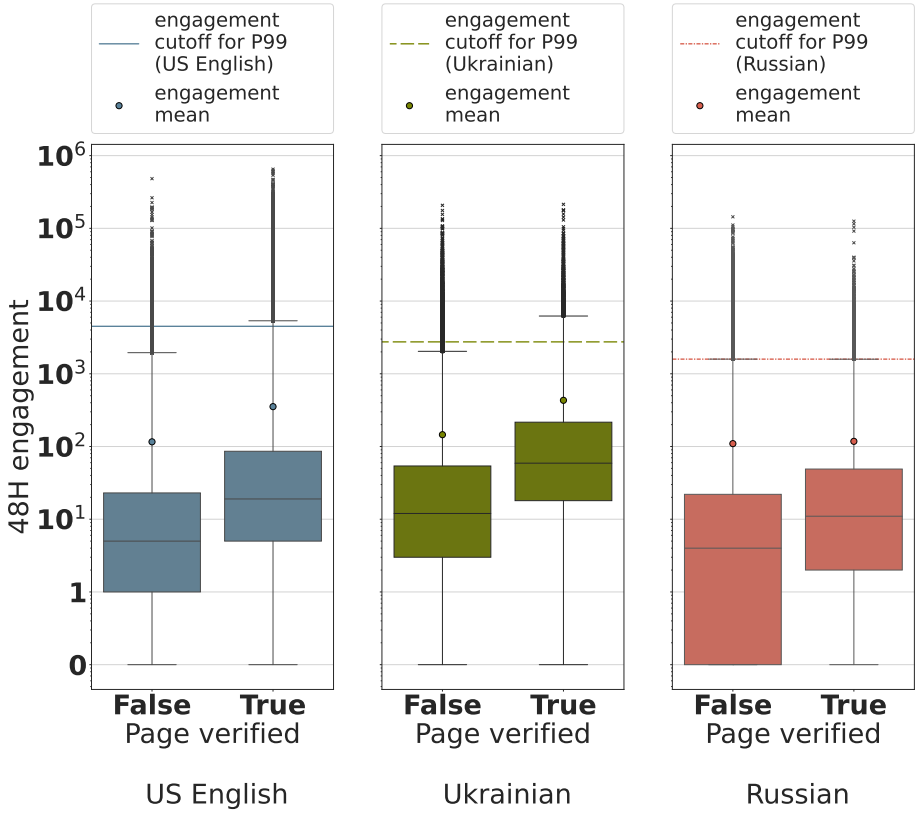

Figure 5: Box plot of distribution of engagement by page verified status (log scale). Whiskers enclose 0–99th percentiles. Verified pages tend to have higher median and mean engagement.

of engagement has already occurred. Median and average half-lives for all classes are shown in Table 4 and demonstrate that in general, posts in higher quartiles of engagement also have longer engagement life cycles.

6 Metrics: Prevented Dissemination and Share of Dissemination Prevented

In Section 7, we will aim to empirically measure how impactful the content removals we observed were on user engagement. This requires a suitable metric. The "content actioned" metric most commonly used in platforms' transparency reports to measure the impact of their own (hard) content moderation systems captures how much policy-violating content exists on the platform. This metric characterizes the status quo of the platform and the end result of content moderation, not how much of a *difference* the moderation efforts made in reducing user exposure to violative content or even whether it was applied promptly. Facebook additionally reports the share of removals that occur before or after a user reports a post (Meta 2024a), and, for some categories, reports a metric it refers to as "prevalence," which it defines as the share of views of content that

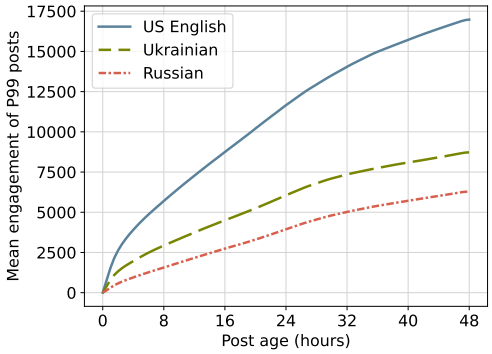

Figure 6: Mean engagement over time for the P99 class of engagement in the US English, Ukrainian, and Russian datasets. P99 posts accrue half of their 48H engagement after an average of 13.6, 12, and 16.2 hours since publication, respectively.

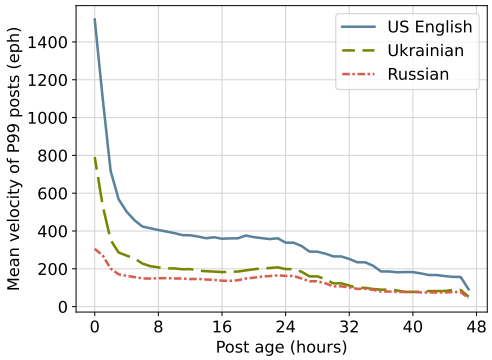

Figure 7: Mean velocity in engagements per hour (EPH) over time since post publication for the P99 class in the US English, Ukrainian, and Russian datasets. Despite their different scale, all three datasets exhibit similar temporal trends. Engagement accrues fastest shortly after posting, before plateauing until a slight pick up in engagement speed occurs around 24 hours after posting.

Table 4: Mean and median half-life engagement values in hours for posts by engagement class. Posts in the US English P99 class reach half of their 48H engagement after an average of 13 hours and 38 minutes. Within each language ecosystem, posts in higher classes of engagement generally (but not always) also had a longer half-life.

Data Set	US English	Ukrainian	Russian
Q1 mean	4h 0m	3h 37m	0h 0m
median	1h 0m	1h 0m	0h 0m
Q2 mean	4h 54m	4h 47m	6h 40m
median	3h 0m	2h 30m	3h 0m
Q3 mean	5h 15m	5h 8m	6h 13m
median	3h 30m	3h 30m	4h 0m
Q4 mean	7h 46m	6h 1m	8h 0m
median	5h 0m	4h 0m	6h 0m
P99 mean	13h 38m	12h 1m	16h 13m
median	11h 30m	9h 30m	16h 30m

violated platform rules (Meta 2019).

We argue that instead, to characterize the effectiveness of a content moderation intervention on the user experience, it is necessary to consider the timing of the intervention in the context of how much dissemination the post would have received had that moderation not occurred. Effectively, we seek to capture two practical realities of content moderation. First, interventions against posts that would otherwise go on to be widely disseminated will have a greater impact on user experience than interventions on similarly violative, but low-dissemination posts. Second, an intervention is most effective when it is applied before anyone has seen a violative post, and least effective after "everyone" has already seen it. We are interested not only in measuring the dissemination that was allowed to happen because of a delay before a violative or borderline post can be identified and an intervention implemented, but also in estimating how much more dissemination the post would have achieved had that intervention not occurred.

We define $D^A_{i,t}$ as the *achieved dissemination* of item i at time t. For this paper, we make the simplifying assumption that an item's dissemination monotonically accrues over time and never decreases (in practice it could decrease as bot-generated views or interactions are removed), and we refer to $D^A_{i,max(T)} = D^{max}_i$ as the item's "final" dissemination at some cutoff time T. This means that at any time t after an item's publication, its total dissemination D^{max}_i can be split into $D^A_{i,t}$, what has already been achieved, and the *previewed dissemination* that is yet to come, which is:

$$D^F_{i,t} = D^{max}_i - D^A_{i,t}$$

We denote $\hat{D^F_{i,t}}$ as the estimate of this measure, and will refer to it as *estimated previewed dissemination*. In practice, we will typically use estimated values for metrics with future time components because of our ultimate goal of measuring prevented dissemination at future times t.

In the content moderation context, items can either be entirely removed from a system, meaning they receive no future dissemination at all (hard moderation), or their dissemination can merely be reduced (soft moderation). We denote *r* as the reduction factor that will be applied to future dissemination after the moderation intervention. In cases where an item is entirely removed, *r* will have a value of 1, and in cases where its dissemination is reduced, *r* will equal the intended degree of reduced dissemination. Therefore, we define our metrics of the *prevented dissemination* of a moderated item *i* at time *t* as:

$$PD^{F}_{i,t} = r * D^{F}_{i,t}$$

and the *share of dissemination prevented* as:

$$PDS_{i,t} = PD^{F}_{i,t} / D^{max}_{i}$$

We denote $\hat{PD}^{F}_{i,t}$ and $\hat{PDS}_{i,t}$ as estimates of these respective metrics, the *estimated prevented dissemination* and the *estimated prevented share of dissemination*.

Dissemination can be measured in a variety of ways. One way of measuring dissemination could be total views or unique users exposed to a post. However, these metrics are not available in our Facebook dataset. As an alternative, in our empirical measurements, we quantify dissemination in terms of *engagement*, which we define as the sum of likes (and other reactions), comments, and shares that a post received. We argue that our methodology can equally be applied to a view-based dissemination metric when such data is available.

7 Case Study: Using Prevented Dissemination to Measure Impact of Content Removal

In this case study, we will first derive a model for each of our three language ecosystems for predicting a post's class of 48H engagement trained on our Primary dataset. Next, we perform an initial analysis of our Removals dataset, and use our model to predict what the 48H class of engagement would have been for these removed posts. Finally, we estimate our metric, *prevented dissemination*, for each removed post, and use this measure to quantify the impact on user engagement of these removal actions.

7.1 Preliminary: Predicting 48H Engagement Classes

In order to assess how impactful post removals were on user engagement (i.e., $\hat{PD}^{F}_{i,t}$ and $\hat{PDS}_{i,t}$), we also need to estimate how much engagement a post would likely have later accrued had it not been removed. In Section 5.2, we identified factors separate from a post itself that are correlated with its engagement, and intuitively, there might be a relationship between past and future engagement. We leverage these insights to develop

a model that predicts a post's future 48H engagement class. While the most intuitive approach to the problem of predicting future engagement is a regression, in practice we were able to achieve better performance by converting this to a classification problem. The extremely large asymmetries of engagement over content and over time mean that we required the ability to tune the model for different portions of the distributions separately, and we also sought the ability to evaluate performance on different parts of the distributions separately.

For each of the three language datasets of non-removed posts, we train a separate multi-class gradient boosting classifier to predict a post's "final" 48H engagement class every hour. Our classifiers use the following features: total post engagement at the time of prediction, velocity at the time of prediction, acceleration at the time of prediction (defined as the rate of change of the two most recent velocity measurements), page subscriber count, weekday of posting, hour of posting, type of post, and page verified status. Each model is trained on the first two weeks of data in the respective language, and tested on the second two weeks of data. For evaluation, we apply the classifier to each post every hour during the 48H window, which is the focus of our study. As time progresses since a post's publication, predictions about the post's ultimate fate gradually improve. Across all languages, our model achieves acceptable accuracy (≥ 74%) predicting posts' class of final engagement at hour 1, and high accuracy (≥ 85%) at hour 7. Table 5 shows the precision and recall for each class and overall accuracy over time.

In addition to understanding the general performance of our model, we would like to know if we are predicting classes accurately before a significant portion of a post's engagement occurs. General measures such as accuracy, precision, and recall treat all posts and all predictions as equally important, obscuring the realities of the asymmetry of engagement we have studied. Our posts are not all equal: some have no engagement at all, while others have tens of thousands of engagements. Also, because our classes are ordinally related, it is more "wrong" to classify a Q1 post as P99 than it is to classify it as Q2. Therefore, when the model predicts a post's class for the next hour, we weight the outcome of the evaluation (correct class or number of classes over/under) by the post's real engagement within the next hour, as observed in the dataset. This weighting by engagement means that we assign a larger penalty for misclassifying a high-engagement post and a lower one for a low-engagement post. We then report the percentage of real next-hour engagement across all posts that corresponded to a correct or over/under class prediction.

We find that for the period between hours 1 and 2, we classify 86.6% of actual engagement correctly, rising to 88.5% of engagement occurring between hours 3 and 4. Over the entire 48H period, we classify 90% of engagement correctly the hour before it happens or earlier, while over-classifying 7.6% by one class and under-classifying 2% by one class. Table 6 shows classification error rates by hour for selected time windows for our

Table 5: Class-of-Engagement model accuracy statistics for US English, Ukrainian, and Russian for selected hours since post publication. Accuracy meets or exceeds 0.9 by hour 12 for all three language ecosystems.

Hour	1	3	6	12	24	36
US English						
Q1 Precision	0.74	0.83	0.85	0.90	0.94	0.98
Recall	0.83	0.88	0.94	0.95	1.0	1.0
Q2 Precision	0.57	0.69	0.78	0.85	0.95	0.98
Recall	0.61	0.74	0.77	0.85	0.91	0.98
Q3 Precision	0.65	0.76	0.82	0.89	0.95	0.99
Recall	0.61	0.74	0.80	0.87	0.94	0.98
Q4 Precision	0.87	0.91	0.93	0.96	0.98	0.99
Recall	0.75	0.83	0.88	0.92	0.97	0.99
P99 Precision	0.80	0.85	0.87	0.89	0.95	0.95
Recall	0.43	0.51	0.65	0.77	0.92	0.91
Overall Accuracy	**0.70**	**0.79**	**0.85**	**0.90**	**0.95**	**0.98**
Ukrainian						
Q1 Precision	0.77	0.84	0.87	0.90	0.94	0.98
Recall	0.82	0.88	0.93	0.95	1.0	1.0
Q2 Precision	0.61	0.73	0.81	0.86	0.95	0.98
Recall	0.65	0.75	0.80	0.86	0.92	0.98
Q3 Precision	0.68	0.79	0.85	0.90	0.96	0.99
Recall	0.66	0.78	0.84	0.89	0.94	0.97
Q4 Precision	0.87	0.92	0.94	0.96	0.98	0.99
Recall	0.80	0.87	0.90	0.93	0.97	0.99
P99 Precision	0.76	0.83	0.83	0.89	0.95	0.94
Recall	0.41	0.55	0.72	0.83	0.93	0.91
Overall Accuracy	**0.73**	**0.82**	**0.87**	**0.91**	**0.96**	**0.98**
Russian						
Q1 Precision	0.65	0.74	0.81	0.88	0.95	0.99
Recall	0.89	0.97	0.99	1.00	1.0	1.0
Q2 Precision	0.56	0.72	0.80	0.88	0.92	0.98
Recall	0.47	0.58	0.70	0.81	0.94	0.99
Q3 Precision	0.66	0.76	0.83	0.88	0.96	0.99
Recall	0.59	0.72	0.79	0.86	0.91	0.98
Q4 Precision	0.86	0.90	0.93	0.95	0.98	0.99
Recall	0.77	0.84	0.89	0.92	0.96	0.98
P99 Precision	0.71	0.79	0.79	0.84	0.94	0.92
Recall	0.32	0.45	0.66	0.76	0.90	0.93
Overall Accuracy	**0.68**	**0.78**	**0.84**	**0.90**	**0.95**	**0.99**

Table 6: Share of engagement underclassified (-*n*), correctly predicted (=), or overclassified (+*n*) by *n* classes across the hour *h* of prediction. Predictions become more accurate as more time elapses since post publication.

h	1	3	6	12	24	36
-4	<0.1%	<0.1%	0%	0%	0%	0%
-3	<0.1%	<0.1%	0%	0%	0%	0%
-2	<0.1%	0%	<0.1%	<0.1%	0%	0%
-1	3.1%	2.2%	2.3%	1.8%	0.9%	0.8%
=	86.6%	88.5%	91%	93.8%	96.8%	97.4%
+1	10%	9.1%	6.6%	4.4%	2.2%	1.8%
+2	0.2%	0.1%	<0.1%	<0.1%	<0.1%	<0.1%
+3	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
+4	<0.1%	<0.1%	0%	0%	0%	0%

Table 7: Mean and median lifetime of removed posts and accrued engagement at the time of last observation.

Data Set	US English	Ukrainian	Russian
Average Last Observation	21h 35m	22h 24m	25h 1m
Median Last Observation	20h 39m	21h 28m	23h 57m
Average Engagement	113	103	65
Median Engagement	5	6	3

US English dataset.

7.2 Dynamics of Content Removals

We now turn to studying how fast content is removed on Facebook, using our Removals dataset. Due to the nature of data collection, we do not know who removed a post, i.e., whether it is an instance of hard content moderation or a voluntary removal.

Overall, removal was a rare event. Only 0.7% of US English posts were removed, compared with 0.2% of Ukrainian posts and 0.5% of Russian. Among our US dataset, average engagement at the time of the last observation was 113, compared to 294 with non-removed data (the latter number is taken at the 48H mark). Comparing medians shows similar effects: 5 at the time of the last observation compared to 14 among our largest set at the 48H mark. Table 7 shows full statistics for the three language ecosystems we study.

In terms of how quickly posts were removed, we observed no strong trend over time, and removals appeared to be nearly uniformly distributed over the observation window for all three language ecosystems. Distribution statistics also reflect this reality; the mean time until the last observation ranged from 21 to 25 hours in our three datasets, with medians only slightly lower and a standard deviation of 11–12 hours. However, given

Table 8: Size of the predicted 48H engagement classes of removed posts. High-engagement classes are underrepresented (i.e., removed posts are estimated to have lower engagement potential compared to non-removed posts).

Dataset	US English	Ukrainian	Russian
Q1	33.6%	41.4%	22.9%
Q2	36.8%	21%	33.5%
Q3	17.4%	19.3%	25.3%
Q4	11.7%	17.6%	18%
P99	0.5%	0.7%	0.3%

that the average engagement half-life of non-removed P99 posts is only 12–16.2 hours depending on the language (and only around 5 hours in the other post classes), this suggests that a majority of posts might already have accrued much of their expected engagement when they were removed.

The question remains to what degree these differences in engagement between removed and non-removed posts are due to the timing of removal or due to the likelihood of removed posts being in lower quartiles of engagement. In order to answer this question and to explore the dynamics of removal between high- and low-engagement posts, we must assign them to engagement classes. However, because these posts were not active at hour 48, we must instead predict their 48H class of engagement, and we use the aforementioned model to do this based on each post's features at the time of the last observation. Because post removals are nearly evenly distributed over the 48-hour time period, the accuracy of these predictions should be considered to be very similar to the prediction model's overall accuracy rate.

Comparing the relative sizes of the predicted 48H classes of removed US English posts, we find that removed posts predicted to be in the bottom two Q1 and Q2 classes are overrepresented (34% and 37%, respectively, as opposed to the expected 25%) and posts predicted to be in the top Q3, Q4, and P99 classes are underrepresented (17%, 12%, and 0.5%, respectively, as opposed to the expected 25%, 24%, and 1%). This means that removals affected mostly lower-engagement posts (in terms of their predicted 48H class), and were disproportionately rare among the most engaging posts. Among removed Ukrainian posts, the predicted 48H Q1 class also appears to be overrepresented, whereas the Russian dataset lacks any obvious bias. Full 48H class prediction results for removed posts in all three languages are shown in Table 8.

We do not find statistically significant differences in removal time between Q1, Q2, Q3, Q4, and P99 engagement classes in any language ecosystem. Figure 8 shows an ECDF of removal times, highlighting the lack of a statistically significant temporal difference between language ecosystems.

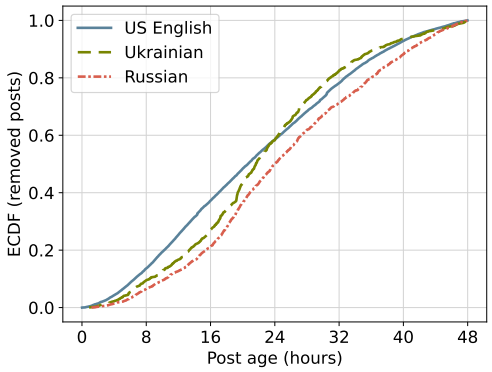

Figure 8: CDF of the age of removed posts at their last observation time for the US English, Ukrainian, and Russian datasets. We visualize all engagement classes for each language together for the sake of readability. We do not find statistically significant differences between either language groups or predicted engagement classes. Removals appear to be nearly uniformly distributed over the 48-hour observation window.

7.3 Results: Estimating Prevented Engagement

Our model predicts classes, but to estimate the impact of removals on user engagement at the time of removal ($\hat{PD}^F_{i,t}$), we first need an estimate of what removed posts' 48H engagement would have been, $D_i^{\hat{max}}$. Recall from Section 7.1 that we evaluate our model accuracy every hour with a test set of the second two weeks of each of our primary datasets. To estimate the value of $D_i^{\hat{max}}$ for a given removed post i removed at time t as accurately as possible, we use the average 48H engagement of the (non-removed) posts from our test set for which we made the same class prediction q at the same post age t. (In doing so, we implicitly assume that the posts in the test set have not been subject to soft moderation, e.g., artificially limited dissemination, which would cause us to underestimate a post's D_i^{max}.) As described in the previous section, we calculate the estimated previewed engagement of a particular post i with post age t at its last observation as $\hat{D}^F_{i,t} = D_i^{\hat{max}} - D^A_{i,t}$. Because these posts were removed entirely, the reduction factor $r = 1$. Thus, prevented engagement of each post i removed at time t can be calculated as:

$$\hat{PD}^F_{i,t} = 1 * \hat{D}^F_{i,t}$$

and the share of 48H engagement prevented by the removal action is:

$$\hat{PDS}_{i,t} = \hat{PD}^F_{i,t} / D_i^{\hat{max}}$$

We perform these calculations for all items in our Removals dataset. We find that in all three languages, most post removals happened long after the posts had already received the majority of their 48H engagement, with removals preventing on average 24.3% of their estimated total 48H engagement in US English, and 30.5% in Ukrainian. Removals of Russian posts occurred late enough in their respective lifecycles that, on average, no engagement was deterred. In US English and Ukrainian, the prevented engagement

is as large as it is because of the extended lifespan of posts in the top 1%. Removals of US English posts in this predicted final class happened on average 23h 40m after publication, but by this point, P99 posts have already received 59% of their total predicted 48H engagement.

While we cannot be certain *why* posts were removed, if the intention was to reduce user exposure to the removed posts, it is clear that the impact in terms of prevented engagement was minor. If these removals were the result of a review pipeline, the share of deterred engagement could have been substantially increased by reviewing the posts earlier. While we estimate that US English and Ukrainian top 1% posts were underrepresented in removals, this class of posts was still responsible for a disproportionate share of actually accrued engagement with removed posts: 55.8% among our US English dataset and 42.7% and 51.5% for our Ukrainian and Russian datasets, respectively. Beyond speeding up removals, identifying more policy-violating P99 posts (if they exist) would have an outsized impact on prevented engagement compared to the other engagement classes.

8 Discussion and Conclusions

Social media platforms are not merely neutral hosts of content. This is, in fact, their primary offering to users: Their recommendation algorithms are capable of surfacing novel content via algorithmic feeds that do not even require users to know what they are looking for. But doing this successfully does not only require platforms to find the best content to surface. Ideally, they should also identify the "worst" content that violates platforms' rules, so they can remove it or otherwise avoid showing it to users. This means that the speed at which recommendation algorithms disseminate content dictates how quickly moderation systems must operate if they are to have an impact on users' experiences.

To understand the temporal environment in which Facebook's moderation currently operates, we first empirically measured the distribution and rate at which content accrues engagement. Similar to prior studies of other platforms (Pfeffer et al. 2023), we discovered news content on Facebook was disseminated quickly, but only had a short period of active engagement. On average, the content we studied received half of their 48H engagement in the first 6 hours after posting. The most engaged 1% of content had longer life cycles, but their engagement was still heavily front-loaded, with half-lives of 12–16.25 hours across the US English, Ukrainian, and Russian news media ecosystems.

Our robust and frequent data collection system (4–12 collections per day) allowed us to observe not only what content was removed, but when it was removed. We cannot say whether the removal events we observed were initiated by the users who posted

those items or by Meta, but we can see that removals typically happen relatively late in posts' engagement life cycles. The average time of last observation for removed posts ranged from 21.5 hours in our US English dataset to 25 hours in our Russian dataset, a longer time period than the post's engagement half-life of even the longest-lived top 1% of posts.

To help measure the impact of content moderation review pipelines in reducing user exposure to policy-violating content, we propose the metric of *prevented dissemination*. This measure effectively captures the value of trade-offs that stem from the two-dimensional asymmetries of attention that we have explored (over different pieces of content and over time), and also can be applied to both hard and soft moderation systems. This is certainly not the only relevant measure of an effective content review system. However, we believe it can contribute a useful view of how impactful different systems can be on the user experience.

By training on our primary dataset of posts and engagement, we were able to create a prototype classifier that can make predictions over time about posts' future engagement based on static features about the posting page, and the dynamic temporal features of currently achieved engagement and velocity of engagement accrual at the time of prediction. Our classifier was able to achieve 90% accuracy about the 48H class of engagement by hour 12 of posts' life cycles. We used this model to estimate the prevented dissemination that resulted from the removal events we observed. We find that overall, these content removals did not have a large impact on how many users engaged with that removed content. Specifically, in the US English and Ukrainian ecosystems, we estimate these removal actions prevented 24% and 30%, respectively, of the engagement those posts would have received. In the Russian ecosystem, the removals happened late enough that no engagement at all appears to have been prevented.

While the removals we observed in the Ukrainian and US English ecosystems did have some effect, if the goal of these removals was to prevent user engagement, there is much room for improvement. Overall, these removals simply happened too late to have a large impact on how many users engaged with the posts. There are two fundamental solutions to this problem: moderation systems can operate faster, or algorithmic feeds can slow down. Our prevented dissemination metric can help inform and assess either of these two approaches.

Unfortunately, both of these two solutions will likely have trade-offs with other platform goals. While some types of content might be interesting to users over a long time period, other posts about natural disasters or breaking news events are most relevant at the time they are posted and quickly become less useful. Slowing down the velocity of feed algorithms will also slow down content people like. Increasing the speed at which moderation pipelines function also has costs. Aside from the obvious monetary costs, aggressively enforcing existing rules also creates the risk of backlash if users perceive rules are being over-enforced.

Our prevented dissemination metric and our findings on the limited impact of content removal actions also highlight a gap in current transparency reporting practices, as well as an opportunity for improvement. Meta does currently report data about views of content later found to violate platform policies, and we would encourage other platforms to do this as well. In addition to being a useful measure for builders of content moderation systems, we believe our prevented dissemination and share of dissemination prevented metrics can also be useful measures for public-facing transparency reporting about moderation systems.

Our empirical findings have several limitations. First, we only study three particular Facebook ecosystems of US-centric public news pages in US English, and media pages in Ukrainian and Russian. It is possible that other ecosystems on Facebook behave differently, and it is known that other platforms have different dissemination characteristics (Lazovich et al. 2022; McGrady et al. 2023). However, the utility of our proposed prevented dissemination metric is not limited to Facebook and could be used with different types of content moderation systems.

Our empirical measurements are also subject to limitations due to a lack of transparency into enforcement actions. The engagement of posts that we measured might have been artificially restricted by "downranking" or other soft enforcement actions. Furthermore, when a post is removed, we do not know whether it was the result of an enforcement action or due to a page voluntarily removing their post. Finally, we can only observe the direct effects of removals that occur after a piece of content is posted. As discussed in Section 4, we cannot observe removals that happen very quickly after a post is created, and we cannot observe indirect consequences of removals, such as other posts that link to identical content being prohibited from being posted in the future.

Our results show that a great deal of user engagement is highly predictable early in a post's lifecycle. Platforms, if they are not already, should be employing predictive tools to prioritize posts in their review pipelines that are likely to have a far reach in the future. Ideally, this should be done as early in the post life cycle as possible, rather than waiting until a post crosses a particular dissemination threshold. If the goal of taking down content is to influence users' experience, then prioritizing resources on a forward-looking basis may be an effective strategy. Given the costs and trade-offs we have discussed, we do not believe there is a single strategy for reducing the dissemination of policy-violating content. In the balancing act between slowing down the dissemination of content that has not been reviewed and reviewing content faster, platforms will have to make choices that take into consideration their goals for the content mix on their platform and their toleration for recommending content that violates their policies to their users.

The period of time in which we undertook this work was marked by a notable retreat from transparency on the part of several social media companies, including Meta. An important future direction of this research would be a deeper analysis of removed content, particularly to better understand the relationship between *when* content is removed and

what is removed. Unfortunately, after the conclusion of our research, Meta also retired CrowdTangle (Meta 2024e). With the removal of CrowdTangle, such research on the Facebook or Instagram platforms has become significantly more difficult, and perhaps out of reach of most independent researchers. Nonetheless, we do not believe that this study would have been possible for independent researchers to perform on any other platform. Meta deserves credit for its (historical) transparency, and we are grateful to the employees of CrowdTangle and Meta for their openness with data that enabled this work.

References

Alhabash, Saleem, and Anna R. McAlister. 2015. "Redefining Virality in Less Broad Strokes: Predicting Viral Behavioral Intentions from Motivations and Uses of Facebook and Twitter." *New Media & Society* 17 (8): 1317–39. https://doi.org/10.1177/1461444814523726.

Almuhimedi, Hazim, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. 2013-02-23. "Tweets Are Forever: A Large-scale Quantitative Analysis of Deleted Tweets." In *Proceedings of the ACM Conference on Computer Supported Cooperative Work,* 897–908. Association for Computing Machinery. https://doi.org/10.1145/2441776.2441878.

Armstrong, Richard A. 2014. "When to use the B onferroni correction." *Ophthalmic and physiological optics* 34 (5): 502–8.

Bhattacharya, Parantapa, and Niloy Ganguly. 2016. "Characterizing Deleted Tweets and Their Authors." In *Proceedings of the International AAAI Conference on Web and Social Media,* 10:547–50. 1. August. https://doi.org/10.1609/icwsm.v10i1.14803.

Bian, Lingyu, Linlin Zhang, Kai Zhao, Hao Wang, and Shengjia Gong. 2021. "Image-based Scam Detection Method Using an Attention Capsule Network." *IEEE Access* 9 (February 16, 2021): 33654–65. https://doi.org/10.1109/ACCESS.2021.3059806.

Buckley, Nicole, and Joseph S. Schafer. 2022. "'Censorship-free' Platforms: Evaluating Content Moderation Policies and Practices of Alternative Social Media." *For(e)Dialogue* 4, no. 1 (February). https://doi.org/10.21428/e3990ae6.483f18da.

Cheng, Justin, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. "Can Cascades Be Predicted?" In *Proceedings of the 23rd International Conference on World Wide Web,* 925–36. Association for Computing Machinery, March 18, 2014. ISBN: 9781450327442. https://doi.org/10.1145/2566486.2567997.

Clegg, Nick. 2023. "New Tools to Support Independent Research." Meta, November. https://about.fb.com/news/2023/11/new-tools-to-support-independent-research/.

Color Of Change. 2020. "Stop Hate For Profit." https://colorofchange.org/stop-hate-for-profit/.

DeVerna, Matthew R., Rachith Aiyappa, Diogo Pacheco, John Bryden, and Filippo Menczer. 2024. "Identifying and Characterizing Superspreaders of Low-Credibility Content on Twitter." *PLOS ONE* 19, no. 5 (May): 1–17. https://doi.org/10.1371/journal.pone.0302201.

Drolsbach, Chiara Patricia, and Nicolas Pröllochs. 2024. "Content Moderation on Social Media in the EU: Insights from the DSA Transparency Database." In *Companion Proceedings of the ACM Web Conference 2024,* 939–42. WWW '24. Singapore, Singapore: Association for Computing Machinery. ISBN: 9798400701726. https://doi.org/10.1145/3589335.3651482.

Edelson, Laura, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. "Understanding Engagement with U.S. (Mis)information News Sources on Facebook." In *IMC '21: Proceedings of the 21st ACM Internet Measurement Conference,* 444–63. Association for Computing Machinery, November 2, 2021. https://doi.org/10.1145/3487552.3487859.

European Commission. 2023. "The Digital Services Act." https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

———. 2024. "DSA Transparency Database." https://transparency.dsa.ec.europa.eu/statement.

Farid, Hany. 2018. "Reining in Online Abuses." *Technology & Innovation* 19, no. 3 (February 9, 2018): 593–99. https://doi.org/10.21300/19.3.2018.593.

Farris, Frank A. 2010. "The Gini Index and Measures of Inequality." *American Mathematical Monthly* 117 (10): 851–64. https://doi.org/10.4169/000298910X523344.

Gastwirth, Joseph L. 1971. "A General Definition of the Lorenz Curve." *Econometrica: Journal of the Econometric Society* (February): 1037–39. https://doi.org/10.2307/1909675.

———. 1972. "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics* 54 (August): 306–16. https://doi.org/10.2307/1937992.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, January. ISBN: 9780300235029. https://doi.org/10.12987/9780300235029.

Goel, Sharad, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2015. "The Structural Virality of Online Diffusion." *Management Science* 62, no. 1 (July 22, 2015). https://doi.org/10.1287/mnsc.2015.2158.

Goodrow, Cristos. 2021. "On YouTube's Recommendation System." *YouTube* (blog), September 15, 2021. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

Google. 2024. "YouTube Community Guidelines Enforcement." Google. https://transparencyreport.google.com/youtube-policy/removals.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7, no. 1 (February 28, 2020): 2053951719897945. https://doi.org/10.31235/osf.io/fj6pg.

Haugen, Frances. n.d. "Summary of Integrity Violations and Solutions." FBarchive. fbarchive.org/user/doc/odoc3656824084w32.

Heath, Alex. 2022. "Facebook is Changing its Algorithm to Take on TikTok, Leaked Memo Reveals." The Verge, last modified June 15, 2022. https://www.theverge.com/2022/6/15/23168887/facebook-discovery-engine-redesign-tiktok.

Heiss, Raffael, Desiree Schmuck, and Jörg Matthes. 2018. "What Drives Interaction in Political Actors' Facebook Posts? Profile and Content Predictors of User Engagement and Political Actors' Reactions." *Information, Communication & Society* 22, no. 10 (March 18, 2018): 1497–513. https://doi.org/10.1080/1369118X.2018.1445273.

Integrity Institute. 2024. *On Risk Assessment and Mitigation for Algorithmic Systems.* Research report. February. https://drive.google.com/file/d/1ZMt7igUcKUq00yakCnbxBCcaA7vajAix/view.

Jenders, Maximilian, Gjergji Kasneci, and Felix Naumann. 2013. "Analyzing and Predicting Viral Tweets." In *Proceedings of the 22nd International Conference on World Wide Web,* 657–64. Association for Computing Machinery, May 13, 2013. https://doi.org/10.1145/2487788.2488017.

Kaushal, Rishabh, Jacob Van De Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. "Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency,* 1121–32. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery. ISBN: 9798400704505. https://doi.org/10.1145/3630106.3658960.

King, Jeff, and Kate Gotimer. 2020. "How We Review Content." Meta, August. https://about.fb.com/news/2020/08/how-we-review-content/.

Lazovich, Tomo, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszar, and Rumman Chowdhury. 2022. "Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics." *Patterns* 3, no. 8 (August 12, 2022). https://doi.org/10.48550/arXiv.2202.01615.

MacCarthy, Mark. 2020. *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry.* Working paper. Transatlantic Working Group on Content Moderation Online and Freedom of Expression, January. https://doi.org/10.2139/ssrn.3615726.

McGrady, Ryan, Kevin Zheng, Rebecca Curran, Jason Baumgartner, and Ethan Zuckerman. 2023. "Dialing for Videos: A Random Sample of YouTube." *Journal of Quantitative Description: Digital Media* 3 (December 20, 2023). https://doi.org/10.51685/jqd.2023.022.

Media Bias/Fact Check. 2021. "Media Bias/Fact Check," February. https://mediabiasfactcheck.com/.

Meta. 2019. “Measuring Prevalence of Violating Content on Facebook.” Meta, May 23, 2019. https://about.fb.com/news/2019/05/measuring-prevalence/.

———. 2023. “About Verified Pages and Profiles.” Meta. https://www.facebook.com/help/196050490547892.

———. 2024a. “Community Standards Enforcement Report.” Meta. https://transparency.fb.com/reports/community-standards-enforcement/.

———. 2024b. “Facebook Community Standards.” Meta. https://transparency.fb.com/policies/community-standards/.

———. 2024c. “How Meta Prioritizes Content for Review.” Meta. https://transparency.fb.com/policies/improving/prioritizing-content-review/.

———. 2024d. “How Technology Detects Violations.” Meta. https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/.

———. 2024e. “Important Update to CrowdTangle.” Meta, March. https://web.archive.org/web/20240425062303/https://help.crowdtangle.com/en/articles/9014544-important-update-to-crowdtangle-march-2024.

———. 2024f. “Our Approach to Facebook Feed Ranking.” Meta, last modified September 4, 2024. https://transparency.meta.com/features/ranking-and-content/.

———. 2024g. “Widely Viewed Content Report: What People See on Facebook.” Meta. https://transparency.fb.com/data/widely-viewed-content-report.

Mohammadinodooshan, Alireza, and Niklas Carlsson. 2023. “Effects of Political Bias and Reliability on Temporal User Engagement with News Articles Shared on Facebook.” In *International Conference on Passive and Active Network Measurement,* 160–87. Springer, March. https://doi.org/10.1007/978-3-031-28486-1_8.

NewsGuard. 2021. “Transparent Tools to Counter Misinformation for Readers, Brands, and Democracies,” March. https://www.newsguardtech.com/.

Papakyriakopoulos, Orestis, Juan Carlos Medina Serrano, and Simon Hegelich. 2020. “The Spread of COVID-19 Conspiracy Theories on Social Media and the Effect of Content Moderation.” *Harvard Kennedy School Misinformation Review* 1, no. 3 (August 18, 2020). https://doi.org/10.37016/mr-2020-034.

Pfeffer, Juergen, Daniel Matter, and Anahit Sargsyan. 2023. “The Half-Life of a Tweet.” *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June): 1163–67. https://doi.org/10.1609/icwsm.v17i1.22228.

Pierri, Francesco, Luca Luceri, Emily Chen, and Emilio Ferrara. 2023. “How Does Twitter Account Moderation Work? Dynamics of Account Creation and Suspension on Twitter During Major Geo Political Events.” *EPJ Data Science* 12, no. 1 (October): 43. ISSN: 2193-1127. https://doi.org/10.1140/epjds/s13688-023-00420-7.

Pierri, Francesco, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. "Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine." In *Proceedings of the 15th ACM Web Science Conference 2023,* 65–74. April 30, 2023. https://doi.org/10.1145/3578503.3583597.

Reddit. 2024. "How Reddit Personalizes Content and Community Recommendations." Reddit, last modified May 16, 2024. https://support.reddithelp.com/hc/en-us/articles/360056999452-How-Reddit-Personalizes-Content-and-Community-Recommendations.

Shiffman, Naomi. 2020. "CrowdTangle API Documentation." GitHub, December. https://github.com/CrowdTangle/API/wiki.

Shiffman, Naomi, and Christina Fan. n.d. "CrowdTangle Codebook." Meta.

Shin, Jieun, Lian Jian, Kevin Driscoll, and François Bar. 2018. "The Diffusion of Misinformation on Social Media: Temporal Pattern, Message, and Source." *Computers in Human Behavior* 83 (June 10, 2018): 278–87. https://doi.org/10.1016/J.CHB.2018.02.008.

Singhal, Mohit, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. "SoK: Content Moderation in Social Media, From Guidelines to Enforcement, and Research to Practice." In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P),* 868–95. IEEE, July. https://doi.org/10.1109/EuroSP57164.2023.00056.

Sorensen, Kiki. 2021. *What's Wrong with Transparency Reporting (and How to Fix It).* Report. December 10, 2021. https://www.adl.org/resources/report/whats-wrong-transparency-reporting-and-how-fix-it.

Sudhakaran, Swathikiran, and Oswald Lanz. 2017. "Learning to Detect Violent Videos Using Convolutional Long Short-term Memory." In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* 1–6. IEEE, October 23, 2017. https://doi.org/10.48550/arXiv.1709.06531.

Taibbi, Matt (@mtaibbi). 2022. "1. Thread: The Twitter Files." Twitter, December 3, 2022. https://twitter.com/mtaibbi/status/1598822959866683394.

Thorgren, Elin, Alireza Mohammadinodooshan, and Niklas Carlsson. 2024. "Temporal Dynamics of User Engagement on Instagram: A Comparative Analysis of Album, Photo, and Video Interactions." In *Proceedings of the 16th ACM Web Science Conference,* 224–34. Association for Computing Machinery, May 21, 2024. https://doi.org/10.1145/3614419.3644029.

TikTok. 2020. "How TikTok Recommends Videos #ForYou." TikTok, June 18, 2020. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.

———. 2024a. "Community Guidelines Enforcement Report." TikTok. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2023-3/.

———. 2024b. “TikTok Community Guidelines.” TikTok. Released April 17, 2024. https://www.tiktok.com/community-guidelines/en/.

Trujillo, Amaury, Tiziano Fagni, and Stefano Cresci. 2024. *The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media.* V. 3, August 1, 2024. https://doi.org/10.1145/3711085. arXiv: cs/2312.10269 [cs.SI].

Twitter. 2023. “Twitter’s Recommendation Algorithm.” *Twitter* (blog), March 31, 2023. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

Vassio, Luca, Michele Garetto, Emilio Leonardi, and Carla Fabiana Chiasserini. 2022. “Mining and Modelling Temporal Dynamics of Followers’ Engagement on Online Social Networks.” *Social Network Analysis and Mining* 12, no. 1 (July 31, 2022): 96. https://doi.org/10.1007/s13278-022-00928-2.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359, no. 6380 (March 9, 2018): 1146–51. https://doi.org/10.1126/science.aap9559.

Wicklin, Rick. 2023. “Weak or Strong? How to Interpret a Spearman or Kendall Correlation.” *SAS* (blog). https://blogs.sas.com/content/iml/2023/04/05/interpret-spearman-kendall-corr.html.

X. 2024. “X Transparency.” X. https://transparency.twitter.com/.

YouTube. 2024. “YouTube Community Guidelines.” Google. https://support.google.com/youtube/answer/9288567?hl=en.

Zhou, Lu, Wenbo Wang, and Keke Chen. 2016. “Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones.” In *Proceedings of the 25th International Conference on World Wide Web,* 603–12. April 11, 2016. https://doi.org/10.1145/2872427.2883052.

Zuckerberg, Mark. 2018. “A Blueprint for Content Governance and Enforcement.” Facebook, November 15, 2018. https://web.archive.org/web/20181117034615/https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

Authors

Laura Edelson is an Assistant Professor at Northeastern University. She can be reached at l.edelson@northeastern.edu.

Borys Kovba holds a Bachelor of Science degree from Igor Sikorsky Kyiv Polytechnic Institute.

Hanna Yershova holds a Bachelor of Science degree from Ukrainian Catholic University.

Austin Botelho is a Research Scientist at New York University.

Damon McCoy is a Professor at New York University.

Tobias Lauinger is a Research Assistant Professor at New York University.

Acknowledgements

We thank Julia Stoyanovich at the Center for Responsible AI at NYU for creating RAI for Ukraine, a fully remote academic research program in partnership with Ukrainian Catholic University in Lviv, Ukraine. We also thank the National Science Foundation, Wellspring Philanthropic Fund, and the Democracy Fund for their support of this work.

Data availability statement

Not applicable.

Funding statement

This work was funded directly by grants from the National Science Foundation (grant number 2344939), the Wellspring Philanthropic Fund, the Democracy Fund, and the Center for Responsible AI at NYU.

Keywords

Social media; internet measurement; content moderation.