RESEARCH-ARTICLE

Causal Insights into Parler's Content Moderation Shift: Effects on Toxicity and Factuality

NIHAL KUMARSWAMY, The University of Texas at Arlington, Arlington, TX, United States

MOHIT SINGHAL, Northeastern University, Boston, MA, United States

SHIRIN NILIZADEH, The University of Texas at Arlington, Arlington, TX, United States

Open Access Support provided by:

Northeastern University

The University of Texas at Arlington

Published: 28 April 2025

Citation in BibTeX format

WWW '25: The ACM Web Conference 2025
April 28 - May 2, 2025
Sydney NSW, Australia

Conference Sponsors:
SIGWEB

Causal Insights into Parler's Content Moderation Shift: Effects on Toxicity and Factuality

Nihal Kumarswamy*
The University of Texas at Arlington
Arlington, Texas, USA
nihal.kumarswamy@mavs.uta.edu

Mohit Singhal†‡*
Northeastern University
Boston, Massachusetts, USA
m.singhal@northeastern.edu

Shirin Nilizadeh
The University of Texas at Arlington
Arlington, Texas, USA
shirin.nilizadeh@uta.edu

Abstract

Social media platforms employ various content moderation techniques to remove harmful, offensive, and toxic content, with moderation levels varying across platforms and evolving over time. Parler, a fringe platform popular among conservative users, initially had minimal moderation, promoting itself as a space for open discussion. However, in 2021, it was removed from the Apple and Google App Stores and suspended from Amazon Web Services due to inadequate moderation of harmful content. After a month-long suspension, Parler returned with stricter guidelines, offering a unique opportunity to study the impact of platform-wide policy changes on user behavior and content outcomes. In this paper, we analyzed Parler data to assess the causal associations of these moderation changes on content toxicity and factuality. Using a longitudinal dataset of 17M posts from 432K users, who were active both before and after replatforming, we employed quasi-experimental analysis, controlling for confounding factors. We introduced a novel approach by using data from another social media platform, Twitter, to account for a critical confounding factor: offline events. This allowed us to isolate the effects of Parler's replatforming policies from external real-world influences. Our findings demonstrate that Parler's moderation changes are causally associated with a significant reduction in all forms of toxicity ($p < 0.001$). Additionally, we observed an increase in the factuality of the news sites shared and a reduction in the number of conspiracy/ pseudoscience sources.

CCS Concepts

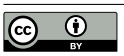
• **Information systems → Web mining**; **Social networking sites**.

Keywords

Parler; Content Moderation Effectiveness; Causal Inference

ACM Reference Format:
Nihal Kumarswamy, Mohit Singhal, and Shirin Nilizadeh. 2025. Causal Insights into Parler's Content Moderation Shift: Effects on Toxicity and Factuality. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714865

*Both authors contributed equally to the paper
†Corresponding author.
‡Work done at The University of Texas at Arlington



WWW '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714865

1 Introduction

Social media has become a powerful tool that reflects the best and worst aspects of human communication. On one hand, they allow individuals to freely express opinions, engage in interpersonal communication, and learn about new trends and stories. On the other hand, they have also become fertile grounds for several forms of abuse, harassment, and the dissemination of misinformation [11, 51, 58, 79]. Social media platforms, hence, continue to adopt and evolve their content moderation techniques and policies to address these issues while trying to respect freedom of speech and promote a healthier online environment.

Social media platforms, however, do not follow unified methods and policies for content moderation [80]. While some social media platforms adopt more stringent content moderation rules, others, like Parler, pursue a laissez-faire approach. Parler, launched in 2018, adhered to this hands-off moderation philosophy, contending that it promoted richer discussions and protected users' freedom of speech [74]. This was until January 6th, 2021, when Parler gained much notoriety for being home to several groups and protesters who stormed Capitol Hill [38, 73]. Subsequently, due to its content moderation policies and concerns about the spread of harmful or extremist content, Parler faced significant consequences. It was not only terminated by its cloud service provider, Amazon Web Services but also removed from major app distribution platforms, including the App Store and the Google Play store [35].

For Paler to return, it had to enact substantial revisions to its *hate speech policies*.[1] This included a complete removal of the ability for users on iOS devices to access objectionable and Not Safe for Work (NSFW) content. As a result, Parler's updated policies introduced more stringent moderation policies aimed at curbing hate speech on the platform [52]. While prior studies have focused on the impact of dealtforming a small subset of users or specific communities [8, 20, 50, 69, 71, 72], or have examined how content moderation affects the activities of problematic users [8, 86], our work evaluates the impact of new platform-wide policy changes on the "within-platform" dynamics of all users who were active during both the pre- and post-policy periods. This inclusive approach offers a comprehensive perspective on the broader ecosystem, moving beyond the limited focus on dealtformed users or specific audiences. Furthermore, while prior research has predominantly examined hard content moderation measures—such as suspensions or removals—our study shifts the focus to *replatforming*, a distinct scenario involving the reinstatement of a platform accompanied by a series of progressive policy changes. In particular, we investigated two research questions:

[1]Example of changes in Parler CG: https://tinyurl.com/yda6pfmj.

RQ1: Did changes to Parler's content moderation guidelines had any significant impact on the user-generated content?

RQ2: How have Parler's content moderation revisions changed its existing users' characteristics?

To assess these effects, we conducted a quasi-experimental analysis, called Difference-in-Difference (DiD) [5], monitoring user posts for toxic content, insults, identity attacks, profanity, and threats. In addition, we explored shifts in users' characteristics and conversation topics, and quantified the presence of biased posts and posts with non-factual links, utilizing data sourced from Media Bias Fact Check (MBFC) [2]. We used the data from Aliapoulios et al. [9] as the seed dataset (we call this dataset a *pre-policy change* dataset), and we tried to collect the posts for the same sample of 4M users. Developing our custom build crawler, we collected about 17M parleys of a subset of 432K users who were active from February 2021 to January 2022. We labeled our dataset as a *post-policy policy change* dataset. To the best of our knowledge, ours is the first dataset that was collected after Parler came back online. To measure the effect of Parler's content moderation changes, we used the Difference-in-Difference (DiD) regression analysis, which is arguably one of the strongest and widely used quasi-experimental methods in causal inference [26, 34, 42, 46]. This analysis helped us understand how and if the outcomes, e.g., the number of toxic posts, have changed after Parler changed its moderation guidelines.

In DiD analysis, to account for the potential influence of offline events, such as social or political unrest, on platform toxicity, we used data from another social media platform—Twitter—as a control group. Our reasoning is that offline events, such as elections, would likely increase toxicity across all platforms, regardless of their content moderation policies. For example, even with strict moderation, Twitter would probably experience heightened toxicity during periods of unrest compared to calmer times. To minimize bias, we analyzed a random sample of Twitter discussions, rather than specifically targeting far-right conversations or incorporating data from other fringe platforms.

Thus, this paper has the following contributions and findings:

(1) Our work demonstrates how the effectiveness of content moderation policies can be evaluated through data-driven analysis using platform-specific data, in this case, Parler, before and after its moderation policy changes.
(2) We collected the first-ever post-replatforming dataset from Parler.
(3) Using the Difference-in-Differences (DiD) approach, we found that Parler was effective in reducing various types of toxicity.
(4) While most related studies fail to account for external offline events that may influence user activity or toxicity, we used trends from a random sample of Twitter data as a control group in our DiD analysis to isolate the effects of the replatforming policies.
(5) Our findings showed an increase in both follower and following counts, along with a rise in users with verified and gold badges. This suggests potential growth in Parler's user base, as well as the continued presence of older users who were active before the moderation policy changes.
(6) We observed an improvement in factuality and credibility scores from the pre-moderation dataset to the post-moderation dataset. Additionally, there was a reduction in the sharing of conspiracy and pseudoscience source links. However, an increase was noted in the sharing of questionable source links in the post-moderation dataset.

2 Related Works

Fringe Communities: Over the past few years, scholars have extensively studied various fringe platforms such as Gab and 4chan [16, 45, 49, 81, 91]. In contrast, Parler is a relatively younger platform, resulting in fewer studies focusing on collecting data or establishing frameworks for data collection from Parler [10, 68]. Some studies have compared topics of discussion on Parler and Twitter [68, 81], most focusing on the presence or prevalence of a single topic. Hitkul et al.[68] examined the Capitol riots—a pivotal event in Parler's history—to compare discussions on Parler and Twitter. Other works have analyzed language use on Parler across various topics, such as QAnon content [13, 81] and COVID-19 vaccines [12]. Our work differs in that we specifically study changes within Parler itself, focusing on how users reacted to the platform's temporary hiatus.

Studies about Deplatforming: All existing studies on deplatforming examine a defined group of deplatformed users and their audiences. Some studies focus on a small number of users (e.g., three users [50]), while others analyze users from specific subreddits (e.g., two subreddits [47, 86] or 15 subreddits [20]) or certain Telegram channels [72]. Since these users are already deplatformed, these studies focus on how their behavior changed after migrating to a new platform. For example, they examine whether activity levels were affected or if there was any change in the use of hate speech [8, 46, 47, 69, 72, 75]. Some studies also track the behavior of audiences across the original and new platforms, investigating questions such as whether followers also migrated to the second platform [69, 72, 75] and, if so, whether they became more extreme in their language [8, 46, 47]. A common finding across these studies is that deplatforming significantly reduces the reach of deplatformed users; however, it also tends to intensify hateful and toxic rhetoric within their new online spaces.

While most of these studies examine user behavior across platforms, only a few focus on "within-platform" dynamics [47, 86]. These works investigate the deplatforming of a small set of subreddits on Reddit, analyzing whether and how the deplatformed users migrated to other subreddits within the same platform. In other words, even when focusing on "within-platform" effects, these studies primarily aim to understand the behavior of a set of deplatformed users and their audiences.

Hate Speech Detection and Classification: Empirical work on toxicity has employed machine learning-based detection algorithms to identify and classify offensive language, hate speech, and cyberbullying [28, 67]. Features including lexical properties, such as n-gram features [60], character n-gram features [57], character n-gram, demographic and geographic features [88], sentiment scores [29, 82], average word and paragraph embeddings [31, 60], and linguistic, psychological, and affective features inferred using an open vocabulary approach [32] have been used to detect hate speech. Google's Perspective API[37] has been widely utilized in prior studies[7, 32, 39, 44, 63, 76, 78, 92] to assess the toxicity of online content. Developed by Jigsaw, the API employs ML models

to assign toxicity scores to text based on various attributes, such as insult, threat, identity attack, and profanity. Despite its widespread adoption, studies have also critiqued its biases, particularly its tendency to over-penalize certain linguistic styles and dialects, raising concerns about fairness and reliability in automated moderation [48].

Media Bias Fact Check (MBFC): MBFC is widely used to assess the credibility and factuality of news sources for downstream analysis [18, 21, 22, 27, 33, 43, 54, 59, 83, 89] and serves as ground truth for prediction tasks [19, 30, 40, 66, 84]. Gruppi et al. [41] used MBFC service to label websites and the tweets pertaining to COVID-19 and 2020 Presidential elections embedded inside these articles. Weld et al. [89] analyzed more than 550M links spanning 4 years on Reddit using MBFC.

3 Methodology

To address our research questions, we first utilized the existing Parler data collected before the policy change [9] and then developed a framework for collecting longitudinal data following the policy change. Second, we utilized Google's perspective API [37] on all the posts to measure various types of toxicity. Third, we analyzed all the links provided in the posts for bias and factuality, using MBFC services. Fourth, we employed Difference-in-difference (DiD) model [5], a quasi-experimental approach, to measure the causal associations of Parler's content moderation change on *toxicity attributes*. In this DiD analysis, we proposed using data from Twitter as a control group to account for a key confounding variable: offline events. Major and contentious events, such as the U.S. presidential election, often drive significant spikes in online activity and can influence toxicity levels across social media platforms. We hypothesize that similar trends in toxicity may be observed on Twitter, which did not implement any policy changes during the same period. By comparing these trends with those observed on Parler, we could assess whether the observed shifts in toxicity levels are unique to Parler's content moderation adjustments or part of broader online dynamics. Therefore, we gathered a sample of Twitter data from the same timeframe and analyzed the trends before and after Parler's policy changes. This approach allows us to isolate the effects of Parler's policy shifts and draw more robust conclusions about their impact on toxicity.

Finally, we examined factors such as the number of followers, following, badge changes, the topics of conversation, and any shifts in the *bias* and *credibility* of the URLs being shared.

3.1 Data Collection

Pre Policy Change Dataset: Aliapoulios et al. [9] developed a data collection tool to gather user information from Parler, capturing data from nearly all active users at the time. The study collected user information from over 13.25M users and randomly selected 4M users for further analysis. For these users, approximately 99M posts (or "parleys") and 85M comments were gathered from August 1, 2018, to January 11, 2021. In our study, we refer to this dataset as the *pre-policy change* dataset. We used these 4M users as a seed dataset to collect data following the policy changes.

Post Policy Change Parler Dataset: We obtained the list of 4M users provided in the pre-policy change dataset for which the authors collected posts and comments [10] and used our custom-build framework to get the content posted by the same users. Using our framework, we collected information about the post body, any URLs posted, a URL to the location of any media posted, the date posted, the number of echoes, and other metadata, such as the badges of the poster. Authors in [9] obtained the metadata of 13.25M users, hence we also tried to collect the metadata of these users.

Post Policy Change Dataset Statistics: From the 4M users, we could collect 17,389,610 parleys from 432,654 active users. Our dataset consists of parleys from February 1st, 2021 to January 15th, 2022. We used the */pages/feed* endpoint, which returns the parleys (posts) posted by a specific user using their username. Note that, this endpoint is different from the endpoint that is used to collect the 13.25M users' metadata, and hence we were only able to obtain 432K users' parleys. Several users from the initial seed dataset of 4M were no longer active. Manually checking these accounts we found that they had either deleted their accounts, or changed their usernames, or did not post any parley after Parler's return, or switched to private accounts. Note, we did not include any users' post if their account was private. We label them as *missing* users. Even though we are unsure if these users were suspended by Parler or they decided to leave Parler, we nevertheless analyzed and compared these users with those that remained active. Since we only collected posts from 432,654 users, we acknowledge that certain trends and analyses conducted might not be accurately reflected on the platform. However, as of January 2022, months after returning to the Apple app store, Parler disclosed that they estimate to have around 700K to 1M active users [3]. This ensures that we have collected a significant part of the data.

These parleys (17M) consisted of users posting around 9M links and plain text in the body. A majority of these posts were primary posts that had no parent. If a parley is an original parley and is not an echo of another parley, it is known as a primary post with no parent. We collected the full-text body, a URL if a link was shared, the title of the parley, the date of creation, flags for trolling, sensitive and self-reported, an upvoted flag, a counter of echoes and likes. We noticed that Parler has a trolling flag, which might be set manually by moderators or automatically by the platform.

We also tried to collect profile information for 13.25M users from the pre-policy change dataset. We used the *pages/profile/view* endpoint, that returns the metadata of the user. We found that 12,497,131 of these users still had a valid Parler account, so we could collect metadata for these users. For a vast majority of the accounts, Parler returned the number of followers, the number of following, status (account available or deleted), the number of and types of badges given to the user, a description of all badges available on Parler at the time of collection, date of parley creation, whether the account is private or public, and also whether the account is being followed by or a follower of the user logged in. A minority of profiles have one or more of these fields missing due to changes on the Parler platform from when the user created the account and the time of data collection.

Twitter Data Collection: For a comparative baseline, we gathered a sample of Twitter data from the same timeframe and analyzed the trends before and after Parler's policy changes. To collect the data, we used Twitter V2 API [87] and collected posts daily using the exact timeline of the datasets. To circumvent the API restrictions,

we collected 27K posts daily and restricted the posts to English. After the data collection was done, we were able to collect 24.16M posts for a pre-policy timeline and 9.69M for a post-policy timeline.

Ethical consideration. We only gathered posts from Parler profiles set to public and did not attempt to access private accounts. We used the same backend APIs that a user browser would request data from. We only obtained the random sample from Twitter API and did not collect any metadata information about the users whose profiles were set as private.

3.2 Measuring Toxicity Scores

We utilized the Google Perspective API, which is a state-of-the-art toxicity detection tool [37]. This AI-based tool investigates the provided text and assigns a score between 0 to 1, with a higher score indicating more severity for a particular attribute. We obtained the following attributes: *Severe toxicity*, *Profanity*, *Identity Attacks*, *Threats*, *Insults*, *Toxicity*. For our study, we collected the likelihood scores for each attribute. While collecting the scores, English was used as the default language for all posts since previous studies showed us that a large majority of Parler's userbase was using English as their language of choice to communicate with other Parler users [9]. Before sending the posts to Perspective API, we pre-processed the posts by removing URLs, hashtags, etc. as these can lead to wrong scores or errors computing the scores by the API. We also pre-processed our Twitter dataset and obtained the scores via Perspective API. While we acknowledge that the Perspective API has limitations in detecting toxicity [48], prior research has demonstrated its effectiveness in identifying various forms of toxic language in generated text [62]. Additionally, several studies have successfully used the Perspective API for toxicity detection [7, 32, 44, 78]. To validate the API's performance, two independent coders labeled 200 randomly selected Parleys. The inter-coder reliability, measured by the Cohen's Kappa score, was 0.7, indicating substantial agreement with the Perspective API's toxicity assessments.

3.3 Causal Inference

We employed a causal inference strategy known as the Difference-in-Differences (DiD) model [5] to measure the impact of Parler's content moderation changes. In our DiD analysis, we assess the causal effect of our dependent variable—toxicity attributes—over time using regression. This is done by comparing two groups: the treatment group (i.e., Parler, where changes in content moderation policies occurred after its return) and the control group (i.e., Twitter, where no such policy changes took place). It is crucial to account for time in the regression; otherwise, we risk misinterpreting a consistent trend (increasing or decreasing) as a treatment effect when simply comparing averages before and after the treatment [15].

To have a balanced dataset, and since our post-moderation dataset spans approximately 11 months, we filtered the pre-moderation dataset for 11 months (i.e., February 2020 to January 2021). We also employed the same step for our Twitter dataset. After filtering the data, we clustered the data points based on the day the parley or the tweet was posted. After clustering the data per day, we set the *Perspective Score* for all the tweets' or parleys' as 0 if they were below 0.5, values above or equal to 0.5 and we kept the absolute value. We choose a threshold of 0.5, because prior research has used this threshold to distinguish if a post is toxic or not [7, 78]. We then averaged the scores per day to get a final score that we passed to our DiD regression model. To check the robustness of our method, we ran a regression model, where we did not use a threshold, and obtained the same results as we obtained when using the threshold, hence our model is robust.

To estimate the effect of the content moderation changes, we employ a linear regression model to estimate the impact (δ) of these policy adjustments following Parler's return:

$$Y = \beta_1 T + \beta_2 P + \delta TP + \epsilon \tag{1}$$

In this model, Y represents the toxicity score; T is a binary variable indicating the treatment group (=1) and control group (=0); and P is a binary variable indicating whether the observation was collected before (=0) or after (=1) the treatment. We estimate the coefficient δ, which corresponds to the interaction between the variables T and P, using Ordinary Least Squares (OLS) to obtain the average treatment effect. β_1 captures the difference between the treatment and control groups prior to the changes in Parler's content moderation guidelines, β_2 reflects the change in the outcome over time for the control group (i.e., Twitter, post-treatment), and δ represents the effect of Parler's content moderation policy changes on toxicity levels.

We chose the DiD method over Interrupted Time Series (ITS) analysis because DiD is widely recognized in the econometrics and causal inference community for handling quasi-experimental interventions [14, 90]. Additionally, DiD provides a single causal estimate (i.e., δ), simplifying the interpretation of results, whereas ITS generates six separate estimates (three for Parler and three for Twitter), making the interpretation more complex.

3.4 Examining Changes in User Characteristics and Content

To answer RQ2 and understand if Parler's moderation change had an impact on its user base beyond users' speech, we performed additional analyses on factors such as the number of followers, following, badge changes, the topics of conversation, and any shifts in the *bias* and *credibility* of the URLs being shared.

3.4.1 User Characteristics. We extracted *following* and *followers* counts from both datasets to understand if any of these metrics have changed significantly after the moderation policy changes. Since these variables are not captured in time, and we have two different distributions, we cannot perform DiD regression analysis. The resulting values did not form a normal distribution, so we used the Mann-Whitney test. We also analyzed the number of badges assigned to each user in the pre- and post-moderation change datasets.

3.4.2 Content Analysis. We used textual data collected from Parleys to investigate what users were discussing in both the pre- and post-policy change datasets. To identify the most popular topics, we applied the Latent Dirichlet Allocation (LDA) topic modeling technique [17]. Before running LDA, we preprocessed the text by removing all URLs, Unicode characters, and stopwords. We also

used a stopword corpus from the Natural Language Toolkit (NLTK) to filter out common words from our dataset.

3.4.3 Assessing Bias and Factuality. We examined the links shared in Parleys to identify trends and align them with the rhetoric of online communities, providing a clearer understanding of the changes. To extract external links, we examined every Parley in both the pre- and post-policy change datasets for valid URLs. We then extracted the top-level domain names from each URL and recorded the frequency of each domain's occurrence to measure website popularity in each dataset. Then, we analyzed these links using the Media Bias Fact Check (MBFC) service [2]. MBFC is an independent organization that uses volunteer and paid contributors to rate and store information about news websites [2]. MBFC can be used to measure the factuality of the URL, the presence of any bias, the country of origin, and the presence of conspiracy or pseudoscience, questionable sources, and pro-science sources. We used a list of links shared from both of our datasets to obtain labels for:

Factuality: Referred to as how factual a website is. Scored between 0-5, where a score of 0 means that a website is not factual and a five is very factual. MBFC defines that for a website to be very factual and get a score of 5, it should pass its fact-checking test as well as make sure that critical information is not omitted.
Bias: MBFC assigns a bias rating of Extreme left, left, left-center, least biased, right-center, right, and extreme right. To assign a bias rating to a website, MBFC contributors check the website's stance on American issues, which divides left-biased websites from right-biased websites [1].
Presence of conspiracy-pseudoscience: Websites that publish unverified information related to known conspiracies such as the New World Order, Illuminati, False flags, aliens, anti-vaccine, etc.
Usage of questionable sources: MBFC defines this as *a questionable source exhibits any of the following: extreme bias, overt propaganda, poor or no sourcing to credible information, a complete lack of transparency, and/or is fake news. Fake News is the deliberate attempt to publish hoaxes and/or disinformation for profit or influence.*

4 Results

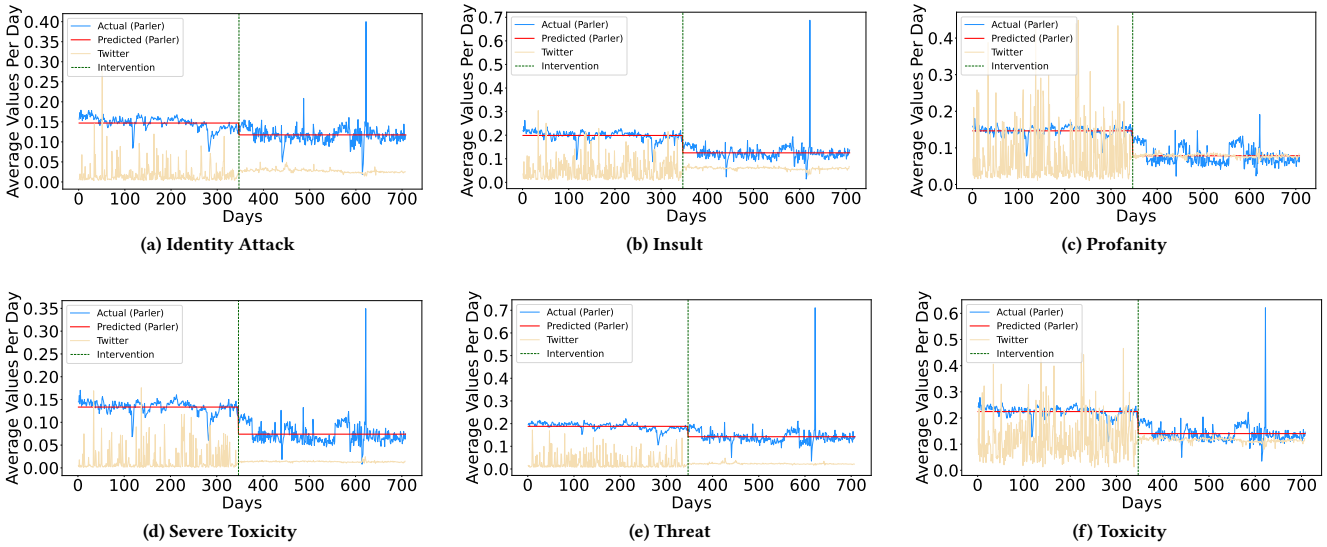
4.1 Impact of Stricter Content Moderation

Table 1 presents the results of our DiD analysis. We observe a causal association between Parler's return online and the implementation of changes to its content moderation guidelines, leading to significant decreases in *Toxicity*, *Severe Toxicity*, *Profanity*, *Threat*, *Insult*, and *Identity Attack* ($p < 0.001$). Additionally, our *Treatment* variable, which indicates whether Parler and Twitter differed in their moderation effectiveness for each dependent variable prior to the policy changes, shows that, on average, Parler had higher levels of *Toxicity*, *Severe Toxicity*, *Profanity*, *Threat*, *Insult*, and *Identity Attack* compared to Twitter ($p < 0.001$). Interestingly, our *Post Treatment* variable shows that Twitter users' *Threat* posts also decreased ($p < 0.05$). This suggests that there could have been a general trend of reduced threatening posts, which could potentially challenge the findings for *Threat* on Parler (δ). However, when examining *Insult* and *Identity Attack*, we observed a statistically significant increase on Twitter ($p < 0.001$), while the opposite trend was observed for Parler (δ). This indicates that despite the general decrease in *Threat* posts, the changes to Parler's content moderation guidelines had a positive effect overall, leading to a reduction in abusive content on the platform. Additionally, Table 1 presents the confidence intervals for all our variables. As observed, for all our *dependent variables*, 95% of the time we would expect the effect of *DiD* (δ) on the various dependent variables to fall within the respective lower and upper bounds. This further demonstrates that the differences in the proposed changes are statistically significant, particularly when examining Parler pre- and post-moderation changes.

Table 1: DiD regression results for toxicity attributes.

Event	*Dependent variable: Toxicity*	Confidence Intervals
Treatment	0.1058 (0.000)***	[0.099, 0.113]
Post Treatment	-0.0028 (0.411)	[-0.010, 0.004]
DiD (δ)	-0.0814 (0.000)***	[-0.091, -0.072]
Event	*Dependent variable: Severe Toxicity*	Confidence Intervals
Treatment	0.1173 (0.000)***	[0.114, 0.120]
Post Treatment	-0.0024 (0.099)	[-0.005, 0.000]
DiD (δ)	-0.0570 (0.000)***	[-0.061, -0.053]
Event	*Dependent variable: Profanity*	Confidence Intervals
Treatment	0.0691 (0.000)***	[0.063, 0.075]
Post Treatment	0.0012 (0.702)	[-0.005, 0.007]
DiD (δ)	-0.0693 (0.000)***	[-0.078, -0.061]
Event	*Dependent variable: Threat*	Confidence Intervals
Treatment	0.1605 (0.000)***	[0.157, 0.164]
Post Treatment	-0.0044 (0.025)*	[-0.008, -0.001]
DiD (δ)	-0.0414 (0.000)***	[-0.047, -0.036]
Event	*Dependent variable: Insult*	Confidence Intervals
Treatment	0.1479 (0.000)***	[0.143, 0.153]
Post Treatment	0.0082 (0.000)***	[0.004, 0.013]
DiD (δ)	-0.0820 (0.000)***	[-0.089, -0.075]
Event	*Dependent variable: Identity Attack*	Confidence Intervals
Treatment	0.1282 (0.000)***	[0.125, 0.131]
Post Treatment	0.0087 (0.000)***	[0.006, 0.012]
DiD (δ)	-0.0382 (0.000)***	[-0.042, -0.034]
Note:	*p<0.05; **p<0.01; ***p<0.001	

To further disentangle these findings, we plotted the results from our DiD model in Figure 1. The red solid line represents δ (i.e., the DiD estimate), while the green line indicates when Parler changed its content moderation guidelines and returned after its hiatus. As observed, across all attributes, the model reveals a statistically significant decrease, with the red line (δ) dropping after the intervention in Parler. Notably, we can visually identify that *Profanity* (Fig.1c), *Severe Toxicity* (Fig.1d), and *Toxicity* (Fig. 1f) exhibit the most significant decreases compared to other attributes. Moreover, we observe that the Twitter pre-intervention time series exhibits much more variability compared to the post-intervention period. We found that this variability was largely due to the difference in the amount of data collected: 24.16 million posts in the pre-policy timeline versus 9.69 million posts in the post-policy timeline dataset. Additionally, we identified fewer posts that were toxic (i.e., above the 0.5 threshold) in the post-policy dataset. We also note a significant peak for all the perspective attributes, except for *Profanity*, around days 600–620. This spike was attributed to the appointment of Jack Smith as the special counsel in the investigations involving Former President Donald Trump [23, 36].



(a) Identity Attack (b) Insult (c) Profanity

(d) Severe Toxicity (e) Threat (f) Toxicity

Figure 1: Difference in Difference (DiD) plots for Perspective Attributes. X-axis denotes the days, and y-axis denotes the average Perspective API scores.

Table 2: Comparison of Users Characteristics

	Pre Policy Change				Post Policy Change			
	Min	Max	Mean	Median	Min	Max	Mean	Median
Followers	0	2,300,000	20.65	1	0	6,048,750	34.8	1
Following	0	126,000	28.28	6	0	479,412	33.4	8

Summary: In summary, our DiD model revealed a statistically significant causal association between Parler's stricter moderation guidelines and a decrease across all toxicity attributes. We observed that the *Treatment* variable (i.e., β_1) was positive for all attributes, indicating that, on average, Parler users posted more abusive content than Twitter users before the intervention. Additionally, *Threat* was the only variable that showed a statistically significant decrease in both Parler and Twitter. Thus, we can conclude that Parler's changes to its content moderation policies had a positive causal effect in decreasing the toxic and abusive content produced by its users, which directly addresses our RQ1.

4.2 User Characteristics and Content

4.2.1 Comparing Users' Characteristics in Pre- and Post-Policy Change Datasets. We employed the Mann-Whitney test and could reject the null hypothesis that users in the pre and post-policy change datasets have the same distribution for *followers* and *followings*. We found an increase in the number of *followings* ($Med_{pre} = 6$ vs. $Med_{post} = 8$) and *followers* ($Mean_{pre} = 20.65$ vs. $Mean_{post} = 34.8$), $p < 0.0001$, hence indicating that users are still active on Parler.

Table 3 presents the number of badges assigned to users in the pre- and post-moderation policy datasets. We observed a significant increase in the number of users undergoing Parler's verification process to confirm that their accounts were not bots. This sharp rise in verified users may reflect concerns about an influx of bots as Parler expanded. Additionally, we found an increase in the number of *Gold* badges, suggesting that some users gained enough popularity post-moderation changes to qualify for this status. These increases indicate that users remained active on the platform following the policy changes. Interestingly, the number of users with the *Private* badge decreased in the post-moderation dataset. Notably, we did not attempt to collect parleys from users with the *Private* badge; rather, badge information was extracted from user metadata.

4.2.2 Content Analysis. We observed significant interest in the 2020 U.S. elections in the pre-policy change dataset, because the elections took place during the data collection period [10]. Additionally, we noticed a decline in the usage of phrases like Where We Go One, We Go All (WWG1WGA), which is associated with the QAnon conspiracy movement. Parler-specific terms, such as Parleys, were also more prevalent in the earlier dataset. Conversely, we observed an increased use of the word patriots, a term Republican lawmakers used to describe the rioters [53].

4.2.3 Links Shared in Parleys. From Table 4, we observe a sharp rise in the popularity of Rumble links (64%). This increase is likely due to Rumble's stance on not removing content related to misinformation and election integrity, with MBFC labeling the website as *Right Biased and Questionable* [56, 80]. In contrast, we observed a decline in the number of Twitter links being shared on Parler. These trends may be attributed to the increasing rhetoric surrounding censorship on Twitter and other popular social media platforms [4]. We also observed a sharp increase in the number of *The Blaze links* (97%) being shared. According to the MBFC service, this website is labeled as *Strongly Right Biased and Questionable* [55].

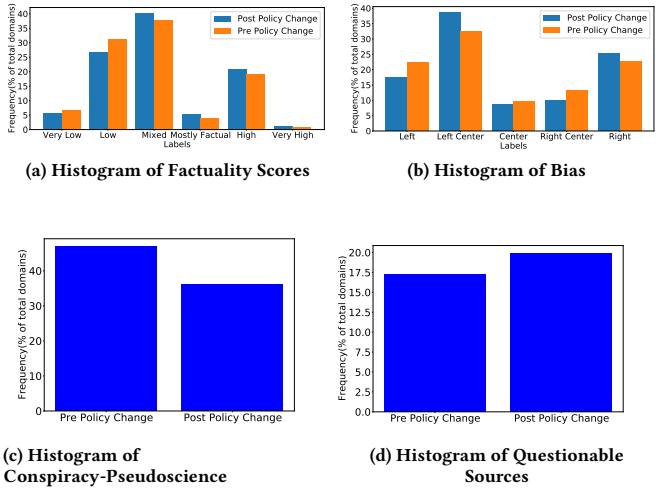
Furthermore, we analyzed the links shared on Parler using the MBFC service. We were able to collect labels for 3,937 (2.59%) and 1,081 (1.75%) of the total links shared on Parleys from the pre- and post-moderation policy change datasets, respectively. However, we could not collect labels for all URLs, as many were from websites,

Table 3: Badges assigned to users in the pre and post-policy policy change datasets

Badge	Description	Pre Change	Post Change
Verified	This badge means Parler has verified the account belongs to a real person and not a bot. Since verified users can change their screen name, the badge does not guarantee one's identity.	25,734	236,431
Gold	A Gold Badge means Parler has verified the identity of the person or organization. Gold Badges can be influencers, public figures, journalists, media outlets, public officials, government entities, businesses, or organizations (including nonprofits). If the account has a Gold Badge, its parleys and comments come from real people.	589	668
Integration Partner	Used by publishers to import articles and other content from their websites	64	N/A
RSS feed	These accounts automatically post articles directly from an outlet's website	99	13
Private	If you see this badge, the account owner has chosen to make the account private. This badge may also be applied to accounts that are locked due to community guideline violations	596,824	337,717
Verified Comments	Users with a verified badge who are restricting comments to only other verified users.	4,147	N/A
Parody	Parler approved parody accounts.	37	N/A
Parler Employee	This badge is applied to Parler employees' personal accounts, should they wish. Their parleys are their own views and not Parler's.	25	28
Real Name	Users using their real name	2	N/A
Parler Early	Signifying Parler's earliest members, this badge appears on accounts opened in 2018.	81	822
Parler Official	These accounts - @Parler, @ParlerDev, and others - issue official statements from the Parler team.	N/A	5

Table 4: Most Popular Websites Shared on Parler

Website	Pre Policy	Post Policy	Change(%)
image-cdn.parler.com	7,318,992	1	−99.99
youtube.com	2,499,198	225,562	−83.44
youtu.be	1,812,871	19	−99.99
bit.ly	893,603	5	−99.99
twitter.com	803,514	42,638	−89.92
media.giphy.com	539,389	545	−99.79
i.imgur.com	532,365	5,779	−97.85
facebook.com	520,796	318	−99.87
thegatewaypundit.com	469,855	610,512	+13.01
breitbart.com	328,953	240,547	−15.52
foxnews.com	298,285	136,956	−37.06
instagram.com	168,160	22,932	−75.99
rumble.com	164,949	744,132	+63.71
theepochtimes.com	136,294	33,937	−60.12
hannity.com	13,017	148,026	+83.83
justthenews.com	50,638	147,984	+49.01
www.theblaze.com	2,006	122,111	+96.76
www.westernjournal.com	6,399	119,551	+89.83
bongino.com	17,251	114,334	+73.77
www.bitchute.com	104,462	87,672	−8.73



(a) Histogram of Factuality Scores

(b) Histogram of Bias

(c) Histogram of Conspiracy-Pseudoscience

(d) Histogram of Questionable Sources

Figure 2: Histogram of MBFC Labels

such as YouTube, Twitter, and Instagram, for which MBFC does not provide labels (see Table 4). Despite this, our results are still generalizable, as we were able to capture the majority of websites for which MBFC does provide labels, allowing us to assess the impact of the policy change on users' speech.

Figure 2 presents our results. We observe a decrease in the number of conspiracy-pseudoscience news articles, as shown in Figure 2c. However, interestingly, there is an increase in the number of *questionable source* articles being shared in the post-policy change dataset, as depicted in Figure 2d. This suggests that Parler continues to allow URLs spreading overt propaganda and fake news, aligning with the findings of [80]. In Figure 2a, we observe that most links with a score between *Very Low* and *Low* are from the pre-policy change dataset, while post-moderation links are more evenly distributed across higher ranges, from *Low* to *High*. Interestingly, in Figure 2b, we observe that Parler users are sharing more URLs from *Left Center* and *Right* websites. This is notable, considering that the majority of Parler users are highly conservative [24]. In summary, using the labels provided by MBFC, we found that the credibility (factuality) of the URLs being shared did increase. Additionally, there was a substantial decrease in the number of conspiracy-pseudoscience news articles. This is particularly interesting, as in 2022, notable conspiracy theories circulated, such as the claims that mpox (monkeypox) was orchestrated by vaccine manufacturers, that Bill Gates was involved in the outbreak, that it was transmitted solely via sexual interactions, and that the WHO released the virus to gain more power [6, 93]. However, despite these shifts, Parler users were now sharing more URLs from questionable sources than before.

Summary: We observed a statistically significant increase in the number of followings after Parler came back online. Additionally, we found that Parler users were more likely to have their accounts *Verified* compared to the pre-moderation change dataset. Interestingly, the term *patriots* became increasingly prevalent in users' posts. We found an increase in the credibility (factuality) of the

URLs being shared. Moreover, we noticed a substantial decrease in the number of conspiracy and pseudoscience news articles. However, Parler users were sharing more questionable source URLs post-moderation than before. These findings indicate substantial changes in Parler users' behavior, effectively answering our RQ2.

5 Discussion and Future Work

Our results indicate a positive impact of the changes to Parler's content moderation guidelines following its ban. Our quasi-experimental analysis revealed that, after the policy changes, all Perspective attributes experienced a statistically significant decrease ($p < 0.001$). As shown in Table 1, we observed that *Severe Toxicity*, *Threat*, and *Identity Attack* saw the largest decreases compared to other attributes. This is particularly interesting as it contrasts with findings from prior studies, which observed an increase in toxic rhetoric among users. In contrast, our research highlights that when Parler adjusted its guidelines, the toxic rhetoric from existing users actually decreased. Additionally, using MBFC, we found that the *credibility* (factuality) of URLs shared by Parler users increased, which directly contradicts the observations made in [86].

Effectiveness of moderation policy. Our results demonstrate that stricter content moderation policies can significantly reduce the toxic rhetoric of existing users. However, an important consideration is the migration of some users from Parler to other fringe social media platforms, such as Rumble, Gab, and Telegram. Studies have shown that users often become more active on fringe platforms after being deplatformed [46], underscoring the need for tailored moderation strategies [25, 80]. Moreover, the migration to fringe platforms can have unintended consequences, with some users exhibiting even more extreme or toxic behavior in these spaces [47]. This raises the question of how content moderation can be effectively managed across multiple platforms, given the risk of users shifting to less regulated environments. Future research should focus on developing coordinated, cross-platform moderation strategies that not only target harmful behavior within individual platforms but also consider the broader ecosystem of social media.

Importance of this study. To the best of our knowledge, this paper is the first study to assess the effectiveness of platform-wide policy changes on all active users present during both the pre- and post-policy periods. This inclusive analysis provides a comprehensive view of the broader ecosystem, as opposed to a limited focus on deplatformed users or specific audiences. Second, while prior research primarily examines hard content moderation measures (e.g., suspensions or removals), our study focuses on replatforming— a unique scenario involving the reinstatement of a platform under a series of progressive policy changes. Third, evaluating the effectiveness of platform-wide moderation is typically challenging due to limited platform transparency and access to comprehensive datasets. However, the unique context of Parler's replatforming allowed us to use a custom crawler to collect data from all active users post-replatforming and to conduct robust comparisons of platform-wide activity and content trends.

Moreover, this study provides the first-ever Parler dataset following its replatforming, along with a framework that can be used to gather data from Parler. This dataset presents a unique opportunity for researchers to examine user behavior, interactions, and the topics discussed within a specific group of users who share particular ideological or political leanings. It offers valuable insights into the dynamics of online communities with more defined mindsets, helping to deepen our understanding of how content moderation and platform policies influence user engagement and discourse.

Furthermore, Parler was acquired by Starboard in 2023 and temporarily shut down on the same day [70]. However, the platform has since returned online, rebranded as Parler 3.0 [65], and implemented significant changes to its content moderation guidelines [64]. As a result, both our dataset and the accompanying framework offer a unique opportunity for researchers to assess the evolution of Parler's policy changes, from its inception in 2018 (Parler 1.0), to the first major policy overhaul in 2021 (Parler 2.0), and now to the current iteration, Parler 3.0. Additionally, Singhal et al.[80] previously found that Parler lacked any form of soft moderation. However, with the introduction of Time Out—a new soft moderation tool under Parler 3.0's updated guidelines[64]—our dataset provides an ideal resource for studying the effectiveness of this intervention in moderating user behavior.

Limitations In our current dataset, i.e., the post-policy change dataset, we were unable to collect a random sample of users, which limits the generalizability of our analysis and may not fully capture the complete impact of the moderation policy change. Another limitation of our study is that some users may have changed their usernames when Parler was reinstated, possibly to evade detection. Additionally, we acknowledge that Google's Perspective API, used for toxicity detection, has certain limitations and biases [61, 77, 85]. Furthermore, our work does not account for the impact of users' hateful rhetoric when they migrated to other platforms after Parler was taken offline. In future research, we plan to investigate user comments on posts to assess whether the moderation changes are reflected in these interactions. Comments may provide additional insights into the effects of the moderation changes Parler implemented.

6 Conclusion

On January 12, 2021, Parler was removed from the Apple and Google App Stores, and Amazon Web Services stopped hosting Parler's content shortly thereafter. This action was attributed to Parler's refusal to remove posts inciting violence following the 2021 U.S. Capitol riots. Parler was later reinstated after strengthening its moderation policies to address hateful content. Our study investigates the impact of these policy changes on user discourse by comparing user rhetoric in pre- and post-policy change datasets. Our quasi-experimental analysis shows that, following the moderation changes, all forms of toxicity experienced a significant decrease ($p < 0.001$). Additionally, we observed an increase in the factuality of the news sites being shared, along with a decrease in the number of conspiracy or pseudoscience sources being shared.

Acknowledgments

This paper is based upon work supported by NSF CNS 2309318 award. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] 2021. Left vs. right bias: How we rate the bias of media sources. https://mediabiasfactcheck.com/left-vs-right-bias-how-we-rate-the-bias-of-media-sources/

[2] 2022. About MBFC. https://mediabiasfactcheck.com/about/

[3] 2022. From the flag-bearer for free speech to 'scapegoat', Parler is fighting back. https://www.thetimes.co.uk/article/from-the-flag-bearer-for-free-speech-to-scapegoat-parler-is-fighting-back-bmwcdfgf5.

[4] Emily A. Vogels, Andrew Perrin, and Monica Anderson. 2020. Most Americans Think Social Media Sites Censor Political Viewpoints. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/

[5] Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *The review of economic studies* 72, 1 (2005), 1–19.

[6] ADL. 2022. Right Wing Lies About Monkeypox Target LGBTQ+ Community. https://www.adl.org/resources/article/right-wing-lies-about-monkeypox-target-lgbtq-community.

[7] Ana Aleksandric, Mohit Singhal, Anne Groggel, and Shirin Nilizadeh. 2022. Understanding the Bystander Effect on Toxic Twitter Conversations. *arXiv preprint arXiv:2211.10764* (2022).

[8] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*. 187–195.

[9] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820* (2021).

[10] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. A Large Open Dataset from the Parler Social Network. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 943–951.

[11] Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2016. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research* 30, 1 (2016), 156–168.

[12] Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. 2021. # Scamdemic,# plandemic, or# scaredemic: What parler social media platform tells us about COVID-19 vaccine. *Vaccines* 9, 5 (2021), 421.

[13] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2023. Finding Qs: Profiling QAnon supporters on Parler. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 34–46.

[14] Allen N Berger and Raluca A Roman. 2020. *TARP and other bank bailouts and bail-ins around the world: Connecting Wall Street, Main Street, and the financial system.* Academic Press.

[15] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology* 46, 1 (2017), 348–355.

[16] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 50–57.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[18] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

[19] Lia Bozarth, Aparajita Saraf, and Ceren Budak. 2020. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 48–59.

[20] Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2024. The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit. (2024), 85–93.

[21] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[22] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific reports* 10, 1 (2020), 1–10.

[23] Zachary Cohen, Scannell Kara, Jeremy Herb, Katelyn Polantz, and Chandelis Duster. 2022. Who is Jack Smith, the special counsel named in the Trump investigations. https://www.cnn.com/2022/11/18/politics/jack-smith-special-counsel/index.html.

[24] Ben Collins. 2021. Increasingly militant 'parler refugees' and Anxious Qanon adherents prep for Doomsday. https://www.nbcnews.com/tech/internet/increasingly-militant-parler-refugees-anxious-qanon-adherents-prep-doomsday-n1254775

[25] Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. 248–251.

[26] William H Crown. 2014. Propensity-score matching in economic analyses: comparison with regression models, instrumental variables, residual inclusion, differences-in-differences, and decomposition methods. *Applied Health Economics and Health Policy* 12 (2014), 7–18.

[27] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What went viral during the 2016 US presidential election. In *International conference on social informatics*. Springer, 143–161.

[28] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[29] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.

[30] Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of YouTube channels using acoustic, textual, and metadata information. *arXiv preprint arXiv:1910.08948* (2019).

[31] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. ACM, 29–30.

[32] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.

[33] Gabriele Etta, Matteo Cinelli, Alessandro Galeazzi, Carlo Michele Valensise, Walter Quattrociocchi, and Mauro Conti. 2022. Comparing the impact of social media regulations on news consumption. *IEEE Transactions on Computational Social Systems* (2022).

[34] Anders Fredriksson and Gustavo Magalhães de Oliveira. 2019. Impact evaluation using Difference-in-Differences. *RAUSP Management Journal* 54 (2019), 519–532.

[35] Brian Fung. 2021. Parler has now been booted by Amazon, Apple and Google | CNN business. https://www.cnn.com/2021/01/09/tech/parler-suspended-apple-app-store/index.html.

[36] Office of Attorney General. 2022. APPOINTMENT OF JOHN L. SMITH AS SPECIAL COUNSEL . https://www.justice.gov/d9/press-releases/attachments/2022/11/18/2022.11.18_order_5559-2022.pdf.

[37] Google Perspective API. 2020. https://www.perspectiveapi.com/.

[38] Lena V Groeger, Jeff Kao, Al Shaw, Moiz Syed, and Maya Eliahou. 2017. What Parler saw during the attack on the Capitol. *Propublica. New York: ProPublica, Inc* (2017).

[39] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.

[40] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2021. NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567* (2021).

[41] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2021. NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. doi:10.48550/ARXIV.2102.04567

[42] Bing Han and Hao Yu. 2019. Causal difference-in-differences estimation for evaluating the impact of semi-continuous medical home scores on health care for children. *Health Services and Outcomes Research Methodology* 19 (2019), 61–78.

[43] Aarash Heydari, Janny Zhang, Shaan Appel, Xinyi Wu, and Gireeja Ranade. 2019. YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos. *arXiv preprint arXiv:1907.00435* (2019).

[44] Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E Smaldino, Goran Muric, and Keith Burghardt. 2023. Auditing elon musk's impact on hate speech and bots. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 1133–1137.

[45] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

[46] Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. 2023. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS nexus* 2, 3 (2023), pgad035.

[47] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[48] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[49] Greta Jasser, Jordan McSwiney, Ed Pertwee, and Savvas Zannettou. 2023. 'Welcome to# GabFam': Far-right virtual community on Gab. *New Media & Society* 25, 7 (2023), 1728–1745.

[50] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[51] Sayeed Ahsan Khan, Mohammed Hazim Alkawaz, and Hewa Majeed Zangana. 2019. The use and abuse of social media for spreading fake news. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. IEEE, 145–148.

[52] Rachel Lerman. 2021. Parler's revamped app will be allowed back on Apple's App Store. https://www.washingtonpost.com/technology/2021/04/19/parler-apple-app-store-reinstate/.

[53] Bess Levin. 2021. REPUBLICAN LAWMAKERS CLAIM JANUARY 6 RIOTERS WERE JUST FRIENDLY GUYS AND GALS TAKING A TOURIST TRIP THROUGH THE CAPITOL. https://www.vanityfair.com/news/2021/05/capitol-attack-tourist-visit.

[54] Thomas J Main. 2018. *The rise of the alt-right.* Brookings Institution Press.

[55] MBFC. 2023. The Blaze – Bias and Credibility. https://mediabiasfactcheck.com/the-blaze/

[56] MBFC. 2023. Rumble – Bias and Credibility. https://mediabiasfactcheck.com/rumble/

[57] Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 299–303.

[58] Trend Micro. 2017. Fake news and Cyber Propaganda: The use and abuse of social media. https://www.trendmicro.com/vinfo/pl/security/news/cybercrime-and-digital-threats/fake-news-cyber-propaganda-the-abuse-of-social-media

[59] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 957–970.

[60] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 145–153.

[61] Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2023. Toxic bias: Perspective api misreads german as more toxic. *arXiv preprint arXiv:2312.12651* (2023).

[62] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1246–1266.

[63] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 885–894.

[64] Parler. 2024. Community Guidelines: User Commitment and Content Standards. https://help.parler.com/en/hc/90853520/40/community-guidelines-user-commitment-and-content-standards?category_id=33.

[65] Parler. 2024. Parler 3.0. https://parler.com/.

[66] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. 2019. A link-based approach to detect media bias in news websites. In *Companion Proceedings of The 2019 World Wide Web Conference*. 742–745.

[67] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433* (2018).

[68] Avinash Prabhu, Dipanwita Guhathakurta, Mallika Subramanian, Manvith Reddy, Shradha Sehgal, Tanvi Karandikar, Amogh Gulati, Udit Arora, Rajiv Ratn Shah, Ponnurangam Kumaraguru, et al. 2021. Capitol (Pat) riots: A comparative study of Twitter and Parler. *arXiv preprint arXiv:2101.06914* (2021).

[69] Adrian Rauchfleisch and Jonas Kaiser. 2021. Deplatforming the far-right: An analysis of YouTube and BitChute. *Available at SSRN* (2021).

[70] Reuters. 2023. https://www.reuters.com/markets/deals/parler-shut-down-temporarily-after-starboard-buys-social-media-platform-2023-04-14/.

[71] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. 2020. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *arXiv preprint arXiv:2010.10397* (2020).

[72] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35, 3 (2020), 213–229.

[73] Candace Rondeaux, Ben Dalton, Cuong Nguyen, Michael Simeone, Thomas Taylor, and Shawn Walker. 2022. Parler and the Road to the Capitol Attack. https://www.newamerica.org/future-frontlines/reports/parler-and-the-road-to-the-capitol-attack/. (2022).

[74] Mike Rothschild. 2021. Parler wants to be the 'free speech' alternative to Twitter. https://www.dailydot.com/debug/what-is-parler-free-speech-social-media-app/

[75] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 742–753.

[76] Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User Engagement and the Toxicity of Tweets. doi:10.48550/ARXIV.2211.03856

[77] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5884–5906. doi:10.18653/v1/2022.naacl-main.431

[78] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*. 1086–1097.

[79] Mohit Singhal, Nihal Kumarswamy, Shreyasi Kinhekar, and Shirin Nilizadeh. 2023. Cybersecurity Misinformation Detection on Social Media: Case Studies on Phishing Reports and Zoom's Threat. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 796–807.

[80] Mohit Singhal, 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)*. IEEE, 868–895.

[81] Andrea Sipka, Aniko Hannak, and Aleksandra Urman. 2022. Comparing the Language of QAnon-related content on Parler, Gab, and Twitter. In *14th ACM Web Science Conference 2022*. 411–421.

[82] Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1481–1490.

[83] Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 230–239.

[84] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 527–537.

[85] Nathan TeBlunthuis, Valerie Hase, and Chung-Hong Chan. 2023. Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *arXiv preprint arXiv:2307.06483* (2023).

[86] Amaury Trujillo and Stefano Cresci. 2022. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW2 (2022), 1–28.

[87] Twitter. 2022. Twitter API. https://developer.twitter.com/en/docs/twitter-api

[88] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[89] Galen Weld, Maria Glenski, and Tim Althoff. 2021. Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. *ICWSM* (2021).

[90] Longqi Yang, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, et al. 2022. The effects of remote work on collaboration among information workers. *Nature human behaviour* 6, 1 (2022), 43–54.

[91] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*. 1007–1014.

[92] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *12TH ACM WEB SCIENCE CONFERENCE*. ACM.

[93] Marco Zenone and Timothy Caulfield. 2022. Using data from a short video social media platform to identify emergent monkeypox conspiracy theories. *JAMA Network Open* 5, 10 (2022), e2236993–e2236993.