Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Political Elites in the Attention Economy: Visibility Over Civility and Credibility?

Ahana Biswas [1], Yu-Ru Lin [1*], Yuehong Cassandra Tai [2], Bruce A. Desmarais [2]

[1] University of Pittsburgh
[2] Pennsylvania State University
ahana.biswas@pitt.edu, yurulin@pitt.edu, yhcasstai@psu.edu, bdesmarais@psu.edu

Abstract

Elected officials have privileged roles in public communication. In contrast to national politicians, whose posting content is more likely to be closely scrutinized by a robust ecosystem of nationally focused media outlets, sub-national politicians are more likely to openly disseminate harmful content with limited media scrutiny. In this paper, we analyze the factors that explain the online visibility of over 6.5K unique state legislators in the US and how their visibility might be impacted by posting low-credibility or uncivil content. We conducted a study of posting on Twitter and Facebook (FB) during 2020-21 to analyze how legislators engage with users on these platforms. The results indicate that distributing content with low-credibility information attracts greater attention from users on FB and Twitter for Republicans. Conversely, posting content that is considered uncivil on Twitter receives less attention. A noticeable scarcity of posts containing uncivil content was observed on FB, which may be attributed to the different communication patterns of legislators on these platforms. In most cases, the effect is more pronounced among the most ideologically extreme legislators. Our research explores the influence exerted by state legislators on online political conversations, with Twitter and FB serving as case studies. Furthermore, it sheds light on the differences in the conduct of political actors on these platforms. This study contributes to a better understanding of the role that political figures play in shaping online political discourse.

1 Introduction

Social media channels are effective for communication due to the ease of information dissemination irrespective of the quality of the information. Politicians also leverage social media due to the wide reach of these platforms. While public officials can promote constructive dialogue, they can also spread harmful content online. Government officials are often subjected to less stringent content moderation rules (Pelletier et al. 2021), and have higher followings than average users (Grant et al. 2010) which makes it easier for them to endorse and propagate harmful content online.

Political communication online is often geared toward the target audience and platform characteristics (Kreiss 2016; Enli and Skogerbø 2013; Stier et al. 2020; Kelm 2020). For instance, Stier et al. (2020) found that social media messages of both candidates and their audiences focused on distinct topics compared to the general audience during 2013 German federal elections. More broadly, politicians often tailor their content to achieve maximum political gains, and prior studies have shown that social media plays a significant role in shaping political agendas, influencing public opinion, and potentially affecting electoral outcomes (Kreiss 2016; Bovet et al. 2019; Boulianne and Larsson 2023).

To understand how politicians use social media to sway public opinion, it is critical to examine what makes them visible online. In this study, we use the term visibility to refer to the level of public engagement or interactions individuals receive on social media, which is often utilized to gauge their outreach and influence on these platforms (Eberl et al. 2020) . Party membership, demographic information (such as state, race, and gender), and platform-level characteristics (such as follower count, posting activity, and content style) are potential contributors. We are particularly interested in determining if politicians' visibility is increased by the dissemination of *harmful* content, which includes toxic and uncivil language, as well as untrustworthy or non-credible information. Both uncivil and low-credibility content have been associated with a decline in the quality of democratic discourse (Goovaerts et al. 2020; Bennett et al. 2018). We ask, *does posting uncivil and low-credibility content increase the visibility of politicians?* The answer to this question is critical as prior research has linked harmful content online to violent offline incidents, increased affective polarization, distrust in institutions, and so on (Johnson 2018; Serrano-Puche 2021; Coe et al. 2014). Harmful content originating from or endorsed by politicians may further exacerbate these negative outcomes due to their larger audience base and higher trustability owing to partisan preferences.

This work focuses on how US state legislators cultivate and exert their influence online. In contrast to national politicians, who are closely scrutinized by media outlets (Kyriakidou et al. 2021), sub-national politicians are more likely to disseminate harmful content with limited scrutiny (Mihailidis et al. 2021). State legislators are responsible for laws across all policy areas within state jurisdiction, making their role crucial in the U.S. political system. Given the limited media coverage of state legislators (Squire et al. 2019), these social media platforms serve as important mediums for com-

*Corresponding author.

municating their ideological and political positions to their voters.

We study the dynamics of legislators' visibility, examining the different factors that influenced the attention they received on Twitter and Facebook (FB) during the two-year period spanning 2020-2021. We focus on these two years owing to the surge of harmful content online due to significant events such as the US Presidential elections, COVID-19, Capitol Riots, and BLM protest movements (Ferrara et al. 2020; Cuan-Baltazar et al. 2020; Toraman et al. 2022). Studying the visibility dynamics over a time period has certain challenges. Apart from individual attributes (e.g., posting frequency, party, demographics) or volume of harmful content posted, the politician's visibility may also vary by time (e.g., during elections) or due to the particular topics they post. We tackle these challenges in our study, our main contributions are as follows :

- **Factors Associated with Visibility.** We present a large-scale, longitudinal study on political elites' online visibility in the US by comparing differences in their platform visibility based on party, socio-demographic factors, and posting activity (RQ1, RQ2; See Section 3) . Republicans and men tend to have a higher level of visibility on FB, while Democrats tend to have higher visibility on Twitter. Posting uncivil content on Twitter and similarly low-credibility content on FB is also correlated with their platform visibility. Legislators' visibility on posting low-credibility content, however, varies by party on Twitter. Our thorough analysis of legislators who post on both platforms reveals notable platform differences associated with their social media activities.
- **Methodological Contribution.** We conduct a causal inference study to examine how legislators' social media posts affect their visibility, particularly when the content is uncivil or less credible (RQ3; See Section 3) . To ensure that our findings are not influenced by potential confounding factors, such as temporal and topical correlations that are common in dynamic text data and can bias the results, we leverage deep learning of potential outcome and matching techniques. Our analytical method helps disambiguate the effect of posting activities.
- **Impact of Harmful Content.** The results, based on observational data using causal inference (RQ3; See Section 3) , have revealed significant and novel patterns. Our study found that posting uncivil content on Twitter led to a decrease in visibility. It was observed that Republicans posting low-credibility content on both platforms have an increased visibility, while Democrats posting the same have lower visibility. The effect is more pronounced for ideologically extreme legislators in most cases. Overall, our findings contribute to the understanding of politicians' online visibility, shedding light on cross-platform differences and partisan asymmetries.

2 Related Work

Political Elites' Online Behaviors. Social media are used by politicians for both broadcasting as well as having dialogue with voters. The effect of Twitter and Facebook use on election campaigns has been studied extensively (Kreiss 2016; Jungherr 2016; Boulianne and Larsson 2023; Sahly et al. 2019). Kreiss (2016) looked at how Twitter was used by political party staffers to shape the perspectives of journalists and influence dedicated voters. Voters engaging in political discussion online have demonstrated increased interest and engagement in political affairs (Bode et al. 2016). Social media, thus, serves as a powerful tool for politicians to influence public opinion and/or convey their stance regarding several important issues. Despite a large body of work on political communication on social media, there is no clear understanding of which factors influence the online visibility of legislators, especially, how politicians posting *harmful* content is viewed by the audience. Our research aims to close this gap by examining the impact of *uncivil* and *low-credibility* content on legislators' visibility by performing a cross-platform study—after accounting for several confounding factors related to their personal attributes, temporal and topical variations.

Misinformation and Virality. There exists a large body of literature characterizing the diffusion of low-credibility content online (Vosoughi et al. 2018; Friggeri et al. 2014; Zollo et al. 2015). Vosoughi et al. (2018) found that falsehoods spread significantly faster, and reached a broader audience as compared to true news on Twitter. Friggeri et al. (2014) found that rumor cascades on FB tend to penetrate deeper into the social network compared to general reshare cascades. Prior works suggest that the sentiment towards misinformation is primarily negative which could be responsible for the variations seen in the diffusion (Vosoughi et al. 2018; Zollo et al. 2015).

A significant body of research has examined the impact of misinformation on the 2016 and 2020 US presidential elections (Bovet et al. 2019; Pennycook and Rand 2021). Prior research has also shown that online misinformation tends to be directed more frequently toward conservative users (Rao et al. 2022; Yang et al. 2023) making them more likely to engage in misinformation. Misinformation may have a higher reach on social media platforms which could be leveraged by politicians to gain visibility, and the extent may vary across ideologies. However, the question of how misinformation originating from public figures is reacted by audiences is less explored. In this work, we aim to illuminate the impact of posting low-credibility content on legislators' visibility.

The Attention Economy and Toxic or Controversial Content. There is no clear consensus on how uncivil content spreads on online platforms (Shmargad et al. 2022; Gervais 2015). Prior works have studied the nature of incivility in online political communication suggesting that engaging in uncivil discourse may have certain benefits for politicians such as political opinion polarization (Bodrunova and Blekanov 2021) or empowerment by voicing criticism against authorities (Bodrunova et al. 2021). Uncivil content was found to be associated with emotionally loaded language which generated strong responses from the audiences (Mutz 2007). Irrespective of the kind of response, this may lead to higher visibility. Coe et al. (2014) found that uncivil comments

on news websites received more negative reactions. Thus, even though engaging in uncivil discourse may have certain political benefits, it is unclear how that is perceived by audiences—a gap that we address in this study.

Confounding with Textual Data. Causal inference with text data is particularly challenging since the assumptions of causal inference (positivity, conditional ignorability, consistency) may not hold when confounding, treatment or outcomes are encoded in text (Feder et al. 2022). For instance, posts on certain topics may be more likely to contain misinformation and also receive higher engagement from the audience. Prior works on extracting confounding from text have utilized unsupervised dimensionality reduction methods (Roberts et al. 2020; Sridhar and Getoor 2019). Recent works leverage neural networks to automatically extract features especially when the confounders in text are not explicitly known (Koch et al. 2021). To address this problem, some works have added transformer layers for text processing to TARNet or Dragonnet (Veitch et al. 2020; Pryzant et al. 2020). We have extended the state-of-the-art techniques to address the confounding factors that are commonly seen in dynamic social media content, such as textual and temporal correlations due to similar topics, events, and personal attributes. To generate content representations, we leverage contextual RoBERTa embeddings (Liu et al. 2019) with other post attributes. We then utilize a fine-tuned Dragonnet model to produce content embeddings that isolate the confounding factors.

3 Study Design

It is crucial to understand how politicians develop influence through online media and factors associated with the influence. This work focuses on state legislators specifically since they may be more likely than national-level politicians to disseminate harmful content owing to limited scrutiny. We ask the following research questions (RQs):

RQ1. How does the legislators' visibility, as measured by the attention they receive, vary based on party affiliation, and individual attributes such as gender, ethnicity, state representation, and social media activity?
RQ2. What attributes of legislators and their posts are associated with their visibility?
RQ3. How does low-credibility or incivility impact the visibility of legislators' posts?

In RQ1 we analyze whether the attention received by legislators varies based on party affiliation, basic demographics, and posting activity. RQ2 examines what characteristics of legislators are most strongly correlated with their online visibility change. RQ1 and RQ2 characterize how visibility varies by legislators' attributes and provide an understanding of factors that may potentially impact their visibility dynamics at the account level. To examine the impact of posting harmful content, RQ3 investigates how low-credibility and incivility *impacts* the attention received by individual posts. More specifically, we study whether low-credibility or incivility increases or decreases a post's visibility where visibility is measured in terms of expected interaction rate. Since the user attention on social media platforms may vary by legislators' attributes (RQ1), and there may be other factors associated with the visibility (RQ2), in RQ3 we estimate the impact of incivility or low-credibility by controlling for these variables as well as temporal and topical variations.

3.1 Datasets

We collect Twitter and FB posts from all US state representatives and senators who held office during 2020-2021 (i.e., each legislator has been in office at least for a certain time between 2020 and 2021) . We focus on Twitter and FB due to the vast amounts[1] of content produced by legislators on these platforms. Of the 8,028 legislators, 5,712 (64%) legislators, comprising 2,943 (61%) Democrats, 2,740 (48.2%) Republicans, and 29 Independents had at least one Twitter account (Kim et al. 2022). For FB, 5,147[2] (64.1%) legislators, comprising 2,215 (48.2%) Democrats, 2,515 (56.0%) Republicans, and 418 other party members, had either an official or a campaign account. We collect all their Twitter and FB posts during 2020-21. It is important to note that many of these accounts were either dormant or inaccessible during our data collection period[3].

Twitter. Our Twitter dataset[4] comprises around 4M tweets posted by 3,551 (44.2%) US state legislators during 2020-2021. The coverage of state legislators ranges from 29.4 to 96.5%, with a mean of 71.6%. This suggests that our dataset comprises a representative legislator population across most US states. We only analyze tweets by Democrats and Republicans due to the insufficient number of posts from Independent legislators (N=29). We calculate the interaction a tweet receives as the sum of likes, retweets, replies, and quotes. Table 1 shows the number of legislators, posts, and median[5] interactions on post. Democrats are more active and receive higher engagement on posts as compared to Republicans. The comparatively higher volume of posts by Democrats indicates that Twitter is a more preferred communication platform for them compared to Republicans. We also construct the intra-legislator follower network, where a directed edge from legislator $A \rightarrow B$ indicates A follows B.

FB. We collect all FB posts[6] by US state legislators during 2020-21. This yields over 493K posts from 5,147 (64.1%) legislators. The coverage of state legislators ranges from 23.5 to 80.6%, with a mean of 57.7%. We similarly focus only on posts from Republicans and Democrats due to a small number of posts from other party members. We use the interaction metric returned by the CrowdTangle API which is the sum of all reactions ('Likes', 'Love', 'Wow', 'Haha', 'Sad', 'Angry', 'Care'), comments, and shares for a post.

[1] https://www.pewresearch.org/internet/2020/07/16/congress-soars-to-new-heights-on-social-media/

[2] FB account information was crawled from Ballotpedia

[3] See Appendix for further details on data collection

[4] Collected using Twitter API v2 before March 2023

[5] The interactions received on posts have a skewed distribution, so we report the median instead of mean.

[6] Collected using CrowdTangle API

	Twitter			FB		
party	#users	#tweets	Int	#users	#posts	Int
Dem	1677	2.25M	8.0	2211	224K	58.0
Rep	1412	889K	6.0	2501	218K	101.0

Table 1: Descriptive statistics for Twitter and FB datasets showing the number of legislators, posts, and median interactions received per post by party.

Table 1 shows the number of legislators, posts, and median interactions on post. The posting frequency is similar for Democrats and Republicans on FB unlike on Twitter. Interestingly, Republicans receive almost double the interaction as compared to Democrats, suggesting that Republicans may have a larger audience base and hence a larger reach on FB. We do not use the follower count for FB data since some of the legislator accounts are official accounts while others are Pages, due to which the follower counts across these different account types vary widely. We are unable to analyze FB's network data due to the lack of access.

3.2 Individuals' Attributes

We characterize legislators based on their platform presence and individual-level characteristics. For Twitter-based attributes, we include their post count, follower count, and in-degree centrality in the intra-legislator follower network. The post count serves as a proxy for measuring how actively the legislators use the platform and follower count indicates the size of their audience base. Both post and follower counts are likely to have an impact on online visibility (Hasan et al. 2022). Legislator's position in their peer network may also impact their visibility (larger follower base, greater potential for content virality, or seniority). To gauge the influence of the legislators among their peers concerning the immediate connections, we leverage the in-degree centrality measure. For FB-based attributes, we include the post count. The individual-level attributes include party affiliation (Republican vs. Democratic), state, gender (Men vs. Women), ethnicity (White vs. Non-White), and ideology scores. We leverage the ideology scores constructed by Shor and McCarty (2011). Around 99.7% (N=3074) and 70.2%[7] (N=3306) of legislators are mapped[8] to their attributes on Twitter and FB, and only those legislators with attributes are analyzed in our study . The final dataset comprises around 62.2% and 53.8% White, and 68.2% and 67.7% men on Twitter and FB respectively, indicating that men and White legislators outnumber women and Non-Whites respectively. There are 2,131 (26.5%) overlapping users (OL) (i.e., legislators having accounts on both Twitter and FB) across Twitter and FB. Figure 8 in Appendix shows the breakdown of ethnicity and gender by party and for OL. Republicans have fewer women users compared to Democrats for both Twitter and FB and both platforms have more Republican men. The representation of Non-White Republicans is higher on FB than on Twitter.

[7]See Data Collection section in Appendix

[8]Ethnicity and gender are mapped using Ballotpedia. Binary genders are used due to insufficient data about non-binary genders.

4 Measuring Posts' Civility, Credibility, and Legislators' Visibility

4.1 Assessing Posts' Civility

We assess the civility of a post based on the toxicity of their language which is a common practice in literature (Frimer et al. 2023; Kim et al. 2021). Incivility in the context of political speech is commonly associated with rudeness according to the study by Stryker et al. (2016). We follow the definition of *toxic* language provided by Google Perspective[9]: *"rude, disrespectful, or unreasonable comment that is likely to make someone leave discussion"*. Based on this definition, it would be reasonable to assume that the toxicity classifier is able to identify the markers of political incivility. We determine the level of toxicity using the "original" model[10] from Detoxify[11] (Hanu and Unitary team 2020). We choose the cutoff for toxicity score as 0.82 based on manual annotation (see Appendix), i.e., posts having a score above 0.82 are considered uncivil[12]. This yields 24,242 (0.8%) and 277 (See Appendix) uncivil posts on Twitter and FB.

Table 2 shows the number of legislators posting uncivil content, posts, and median interactions received, by party. This may be attributed to politicians' different communication styles across these platforms as observed in prior research, suggesting FB is used more for broadcasting purposes compared to Twitter which is leveraged more for having dialogue (Enli and Skogerbø 2013). For our analysis on uncivil content in the following sections, we only focus on Twitter owing to the small number of uncivil posts on FB. Around 47.1% of Democrats and 32.6% Republican legislators post uncivil content on Twitter. Democrats post almost double the number of uncivil content compared to Republicans (on average) on Twitter. This could be due to the higher interaction received on such posts by Democratic legislators. The interaction on uncivil content is higher compared to the baseline interaction (in Table 1) for both parties, with a larger difference for Democrats.

Figure 1A shows the rate of uncivil posts across years by party and platform. More uncivil content was posted during 2020 on Twitter, with Democrats having a higher rate of posting uncivil content. Figure 2A shows the rate of uncivil posts across states by party and by platform. Interestingly, we find that Republicans posted more uncivil content on FB compared to Democrats across almost all the states, but state-wise differences exist for Twitter.

4.2 Assessing Posts' Credibility

We identify low-credibility content based on the credibility of URL in the post, which is a common practice in liter-

[9]https://perspectiveapi.com/how-it-works/

[10]The "original" model had the best performance when evaluated against our manual annotation labels compared to the other Detoxify models, namely, "debiased" and "multilingual".

[11]We choose Detoxify over Google Perspective API since Detoxify has better or comparable performance (Arhin et al. 2021) and is faster

[12]See Appendix for examples of uncivil posts

	Twitter			FB		
party	#users	#tweets (%)	Int	#users	#posts (%)	Int
Dem	782	18111 (0.9%)	14.0	55	89 (0.1%)	131
Rep	461	6131 (0.7%)	7.0	54	188 (0.1%)	120

Table 2: Number of uncivil posts, legislators posting, and median interactions received per post on uncivil content on Twitter and FB, by party.

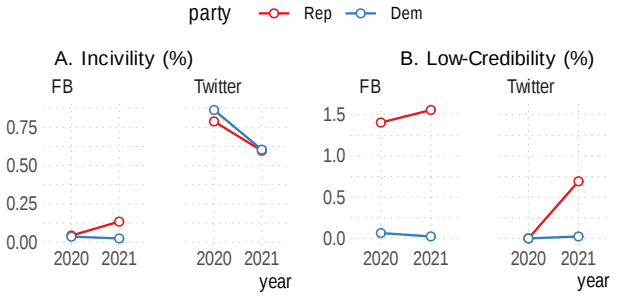

Figure 1: **Percentage of uncivil and low-credibility posts across years, by party, by platform.** Republicans have a higher rate of posting low-credibility content on both platforms and across years. The yearly posting rates of uncivil content are comparable across parties.

ature (Lasser et al. 2022). It is important to note that this method of identifying low-credibility posts does not discriminate between posts endorsing or debunking such information. In this study, we are interested in looking at the visibility of both, i.e., posts promoting or debunking low-credibility information since both types of posts are exposing the audience to harmful information. Prior research has shown that exposure to misinformation has an impact on believing and subsequent sharing of such information (Halpern et al. 2019). Thus, irrespective of the author's intent, the audience may be susceptible to believing and/or sharing such low-credibility content. Moreover, the URL sharing patterns are similar for Democrats and Republicans on both Twitter and FB[13], so it is unlikely that any biases are introduced in the study due to our method of using URLs to identify low-credibility posts. We use the low-fact URL domain references provided by Tai et al. (2023) to identify low-credibility posts. Tai et al. (2023) refined Media Bias/Fact Check's[14] (MBFC) original ratings to include URLs that contain misleading information and not just politically biased ones. This restrictive framing may undermine the true scale of misinformation on these platforms but it offers a higher precision (vs. recall) in identifying low-credibility content. This yields 6,848 (0.2%) and 4,141 (1.0%) low-credibility posts on Twitter and FB respectively suggesting that legislators post more low-credibility content on FB.

[13]Around 19.2% posts by Republicans and 17.4% by Democrats contain URLs on Twitter, and 14.6% and 18.3% posts by Republicans and Democrats on FB contain URLs.

[14]https://mediabiasfactcheck.com/

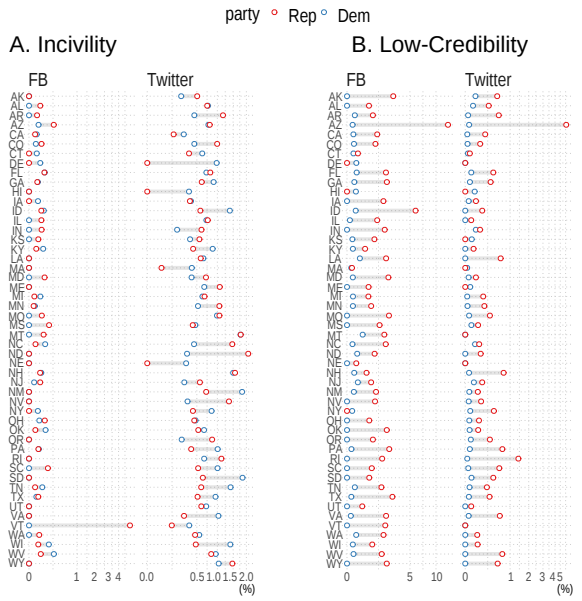

Figure 2: **Percentage of uncivil and low-credibility posts from states, by party, by platform.** Republicans have a higher rate of posting low-credibility content across all states, on both platforms. Moreover, Republican legislators from most states have a higher rate of posting uncivil content on FB. State and party-wise differences exist for posting rates of uncivil content on Twitter. Note that the x-axis scale has been adjusted to better visualize party differences.

Table 3 shows the number of legislators posting low-credibility content, posts, and median interactions received, by party. Around 5.2% of Democrats and 36.7% Republican legislators post low-credibility on FB. On Twitter, only a handful of accounts are responsible for spreading low-credibility content across both parties (1.1% for Democrats and 2.8% for Republicans). Republicans post more low-credibility content compared to Democrats on both platforms. Similar to uncivil content, posts containing low-fact URLs receive higher interaction except for Democrats on Twitter. The median interaction for low-credibility content is almost three times on Twitter and double on FB for Republicans compared to their baseline interaction[15].

Figure 1B shows the rate of low-credibility posts across years by party and platform. The prevalence was higher during 2021 on both platforms and it was mainly driven by Republicans. Figure 2B shows the rate of low-credibility posts across states by party and by platform. Low-credibility content is driven by Republicans across all the states, with the highest rate from Arizona, on both Twitter and FB. The low-

[15]For Twitter, the median interaction on posts with and without URLs are 9.0 and 7.0 respectively whereas for FB, the median interactions are 62.0 and 81.0, i.e., there is no clear pattern as to whether having a URL increases or decreases the interaction of posts, suggesting that the results in Table 3 could potentially be due to the low-credibility of URLs.

	Twitter			FB		
party	**#users**	**#tweets (%)**	**Int**	**#users**	**#posts (%)**	**Int**
Dem	19	188 (0.0%)	4.0	83	114 (0.1%)	78.5
Rep	40	6660 (0.8%)	20.0	567	4027 (2.6%)	239.0

Table 3: Number of posts containing low-fact URLs, legislators posting, and median interactions received per post on low-credibility content on Twitter and FB, by party.

credibility information from Arizona is mostly related to the 2020 US Presidential elections. In particular, we find that some Republican legislators frequently shared posts from unreliable information outlets, especially during Arizona's audit of the 2020 election, which contributed to a significant amount of low-credibility posts from Arizona.

The differences in interactions received by low-credibility and uncivil posts may be attributed to the differences in post topics, timing, or attributes of authors—we address this in the following sections.

4.3 Measuring Legislator's Visibility

Social media is being increasingly used as a tool by politicians to enhance their visibility and outreach to the public (Bahramirad 2022). The measure of a legislator's visibility is typically based on the interactions received on their posts. Our approach to measuring visibility is inspired by the metrics used on Twitter and Facebook to calculate a post's expected engagement or virality potential (Twitter-team 2024; Crowdtangle 2024). Our objective is to measure the overall outreach of a post or legislator on the platform. Therefore, we consider all the interactions received on a post or by legislators instead of focusing on a single type of engagement such as "Likes." It's important to note that the visibility metrics are platform-specific and may not be comparable across platforms even if they share the same names. For instance, audiences may engage with "Like" feature on Twitter differently than "Like" on Facebook due to different interface designs (see details in Appendix). However, our visibility metrics are designed to capture the overall level of engagement a post or legislator receives on a specific platform.

Moreover, we do not distinguish between positive and negative reactions received from the audience (e.g., 'love' vs. 'angry' on FB). We are concerned with the visibility of the legislators and both positive and negative reactions contribute to their overall outreach on the platform. As shown in Tables 2 and 3, the interactions received on low-credibility and uncivil content are noticeably different from the overall interactions received by legislators, suggesting that audiences' reactions differ based on content (or other related factors). Thus, using interactions as a proxy to measure legislators' online visibility allows us to capture these differences and get insights into factors contributing to their visibility.

The visibility can be measured at both account and post level. For post level, the visibility of i^{th} post by legislator u, v_{ui} is simply the interaction received on that post. Furthermore, we examine how visibility *changes* in relation to other variables. Instead of measuring aggregated interactions over the audience of their posts, we measure the legislator's visibility (V) by interaction rate, i.e., the number of interactions per post or audience size. We consider three different ways to adjust the quantity of the interaction rate, since platform reach tends to be correlated with the number of followers (Hasan et al. 2022), resulting in the following *dependent variables* (DVs): (1) interaction normalized by post count (V^{I_P}), (2) interaction normalized by follower count (V^{I_F}), and (3) interaction normalized by follower and post count ($V^{I_{P,F}}$). Therefore, the visibility of legislator u, V_u, is given by aggregating the visibility of all u's posts normalized by post and/or follower count.

5 Methods

In this section, we describe the methods used to answer our RQs. In addition to the legislator behavior on Twitter and FB, we analyze the overlapping legislators (OL) to understand whether the differences observed across these platforms are due to different legislator populations or different audience/platform characteristics on Twitter and FB.

5.1 Analyzing Legislators' Visibility

In RQ1, we analyze whether legislators' visibility varies based on demographics, party, and posting activity. For attention disparity based on posting activity, we compare visibility of legislators having above the median number of posts with those posting less than or equal to median[16]. Using Mann-Whitney U test, we find differences in the platform visibility of legislators based on party, gender, ethnicity, and posting frequency. We incorporate additional DV variants, namely, $25th$, $50th$, and $75th$ percentile of the legislator's post interactions together with the mean interaction, i.e., V^{I_P} to ensure robustness of our results. For states-wise differences, we visualize the mean visibility across states.

For RQ2, we study the factors significantly correlated with the platform visibility of legislators. In addition to the individuals' attributes (Section 3.2), we measure the association between low-credibility and uncivil post volumes and their visibility. Since low-credibility content is targeted more towards conservative users (Rao et al. 2022; Yang et al. 2023), we also consider the interaction between party and low-credibility content posted. Additionally, the legislator's visibility may be influenced by the visibility of their peers if their peers (re)post similar content often. To measure this network visibility, we use the median of the visibilities of legislator's in-degree neighbors in the intra-legislator follower network. Unfortunately, we are unable to account for the network effects on FB. We use the following model,

$$\begin{aligned} V_u = \beta_0 + \beta_P Party_u + \beta_G Gender_u + \beta_E Ethnicity_u + \\ \beta_N Posts_u + \beta_F Followers_u + \beta_C Centrality_u + \\ \beta_S State_u + \beta_T Uncivil_u + \beta_M LowCredible_u + \\ \beta_{Net} NV_u + \beta_e Party_u * LowCredible_u \end{aligned} \tag{1}$$

where V_u is the visibility of legislator u, $Uncivil_u$ and $LowCredible_u$ are the count of uncivil and

[16]We chose median as the threshold due to the skewed posting activity distribution.

low-credibility posts, $Party_u, Gender_u, Ethnicity_u$ and $State_u$ are dummy variables, $Posts_u$ is the post count, $Followers_u$ is the follower count[17] (Twitter specific), $Centrality_u$ is the indegree centrality in intra-legislator follower network (Twitter specific), and NV_u is the network visibility. To address the correlated errors across and within states, we incorporate a random effect on the state variable. The effect of each factor is estimated using a linear mixed effects regression model[18] with standardized continuous variables. To ensure robust results for the analysis of RQ1 and RQ2, we conduct separate analyses for the years 2020 and 2021, to determine if the visibility trends observed were consistent across both years[19]. Moreover, we analyze the topical[20] distributions (e.g., COVID-19, BLM, elections) for our dataset and find that our data is not skewed towards any particular topic, suggesting limited bias due to specific topic(s).

5.2 Analyzing the Impact of Low-credibility or Uncivil Content

Measuring Outcome. Posting low-credibility and uncivil content may have an impact on how politicians are perceived online. In particular, we analyze whether the presence of incivility or low-fact URLs increases/decreases their posts' visibility. To characterize the change in visibility, we analyze the engagement on a post considering the post author's expected visibility. Our metric is inspired by the overperforming metric used at CrowdTangle (Crowdtangle 2024). The overperforming score for post i is calculated as follows,

$$O_{ui_p} = \frac{v_{ui_p}}{b_{u_p} + thres_p} \tag{2}$$

where b_u is the median interaction for legislator u's posts on the platform in previous w-days and a threshold ($thres$) for the minimum number of interactions on a post to be considered as overperforming, with p denoting platform. The term b_{u_p} is used to adjust the outcome with respect to the level of expected interactions with the post authors on platform p. We choose the $thres = 10$ for Twitter (i.e., a post must have at least 10 interactions to be considered overperforming on Twitter) and 100 for FB. We estimate the ideal window w, based on legislators' daily posting rates on these platforms (see Appendix for $thres$, w estimation). To get a reliable estimate of the overperforming metric, we choose $w = 14$ day rolling window. This allows us to calculate the overperforming score over a sufficient number of posts per legislator, while also accounting for the temporal variation.

Figure 3 shows the post overperforming scores on Twitter and FB[21]. The distribution of overperforming scores are similar for Republicans and Democrats on Twitter. On FB, posts by Republican legislators overperform more compared to posts by Democrats. This suggests that posts by Republican legislators have a higher tendency to be viral on FB compared to Democrats. For estimating the causal impact, we are interested in analyzing whether a post overperforms or not due to its incivility or low-credibility, hence we binarize the overperforming score (*outcome*) at cutoff 1, i.e., a post is overperforming if it has a score > 1.

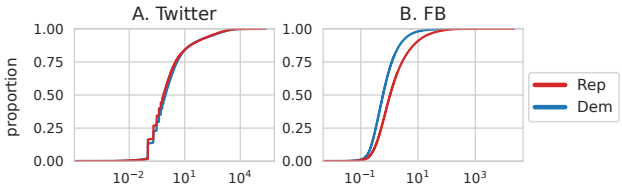

Figure 3: **ECDF plots of overperforming score distributions by party, by platform.** Distributions are similar for Twitter but posts by Republicans overperform more on FB.

Estimating Causal Impact. It is necessary to control for the *confounders* that affect both the *treatment* and *outcome* variables to differentiate between *spurious correlation* and *causation*. One of the possible confounders is the topic—posts on certain topics (e.g., COVID-19) may be more likely to contain misinformation, and these topics may get higher visibility. Another possible confounder could be the tone—posts having certain tones (e.g., assertive) may be more likely to contain uncivil language, and also more likely to generate stronger responses from the audience. Apart from textual content, other confounding variables can include legislators' personal traits and the timing of their posts (e.g., during elections there might be a rise in harmful content as well as an increase in legislators' visibility).

We control for the individual's attributes mentioned in Section 3.2 (excluding ideological scores), post content, and time, i.e., count of days since the content was posted, starting from 2020-01-01. To control for the content, we leverage embeddings generated by the pre-trained RoBERTa model. For low-credibility content, we also include the URL headlines along with post content because we assume that both the text and news headlines are visible to the viewers. The textual embeddings[22] are concatenated with individual's attributes, and time variable to get the final embedding of each post. The confounders we choose to control for are based on prior literature (Hasan et al. 2022; Sahly et al. 2019) and their feasibility of being measured. Apart from these confounders, there could also be certain other confounders (e.g., external events, algorithmic promotion effects, effect of ads) which we are unable to measure and thus account for in this

[17] Only included for DV not normalized by follower count

[18] To satisfy assumptions of linear regression, all variables are suitably transformed to be close to normal distributions (See Appendix). Ideology score is dropped since it is correlated with party.

[19] Our findings revealed that the trends were similar for both years. Therefore, we have reported the overall results for the entire study period.

[20] Identified using keywords.

[21] Our scores for FB posts are highly correlated (Spearman rho=0.997, $p < 0.001$) with the overperforming score returned by the CrowdTangle API, suggesting that our method successfully identifies posts that are overperforming.

[22] Only posts having a minimum of 10 words are considered for this part of the analysis. This accounts for 5,405 (78.9%) low-credibility and 15,883 (65.5%) uncivil posts on Twitter, and 2,427 (58.6%) low-credibility posts on FB.

study. Our causal effect estimation has two steps: potential outcome prediction and matching.

Potential outcome prediction. The confounding in our case is time-varying and encoded in complex textual data, so we leverage the non-parametric nature of neural networks to learn deconfounded, low-dimensional representations of the high-dimensional data (Koch et al. 2021). We leverage the Dragonnet[23] model proposed by Shi et al. (2019) which uses feed-forward neural networks to learn balanced[24] representations of the data such that each head models a separate potential outcome, a third propensity head predicts the propensity (π) of being treated and a free nudge parameter ϵ (see Appendix for model description).

We fine-tune the Dragonnet model[25] by adding more hidden layers and using cross-entropy loss. The *treatment variables* in our case are low-credibility and incivility respectively. We use a 5-fold cross-validation setting for training, with a 1:1 ratio of treated vs. non-treated random samples (see Appendix for model performance). For low-credibility posts, we select corresponding non-treated posts that contain at least one URL and similarly include the URL headlines to minimize the confounding from text (e.g., presence of URL). Figure 4 shows the effectiveness of our deconfounding for incivility on Twitter.

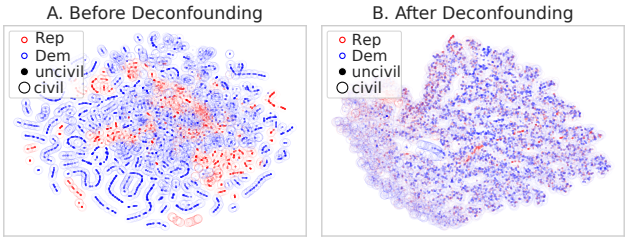

Figure 4: **Effectiveness of deconfounding for uncivil vs. civil tweets.** (A) shows the T-SNE space of content embeddings for civil vs. uncivil tweets by party. (B) shows the representation of the deconfounded embeddings returned by Dragonnet. After deconfounding, the representation space for treated and control covariate distributions (*party* in this example) can not be distinguished.

Matching. For *Conditional Average Treatment Effect* (CATE) estimation with Dragonnet predictions, we find that the covariates are not balanced after propensity score reweighting. To improve balance, we further use matching. We match the treated and untreated subjects based on the deconfounded Dragonnet embeddings (see Appendix). The covariate balance for matching is shown in Figure 5. All the covariates are balanced for Twitter and FB low-credibility

[23]Dragonnet is chosen over other deep learning models for causal inference (S-learner, T-learner, TARNet) due to its "Targeted Regularization" procedure which allows for statistical guarantees (Koch et al. 2021)

[24]Balancing is a treatment adjustment strategy that forces the treated and non-treated covariate distributions closer to deconfound the treatment from outcome (Johansson et al. 2016)

[25]Models trained on a single NVIDIA A100 40GB PCIe GPU

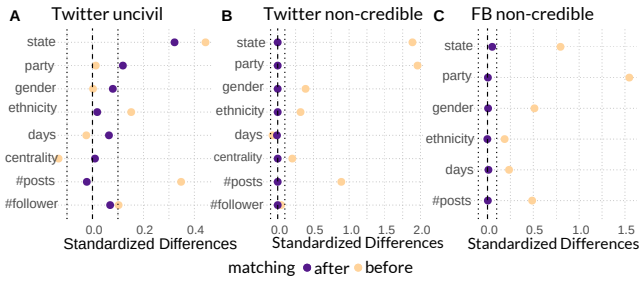

Figure 5: **Covariate balance after matching.** All covariates, except for party, and state in Twitter uncivil model, are balanced (i.e., absolute standardized difference < 0.1) after matching on deconfounded embeddings.

models. For Twitter incivility, all the covariates except for state and party are balanced after matching. The final CATE is calculated based on the matched samples as follows,

$$CATE = \frac{1}{N'} \sum^{N'} (Y_i(1) - Y_{i'}(0)) \quad (3)$$

where $Y(T)$ is the outcome for treatment T and N' is the number of matched samples. We estimate the CATE for Democrats and Republicans separately to study potential asymmetries in receptivity across their audiences. Furthermore, we look at the CATE for ideologically extreme legislators to understand whether audiences engage differently with legislators at the extreme. We consider legislators having top 25% conservative and top 25% liberal ideological scores as Extreme Republicans and Extreme Democrats.

Furthermore, our analysis ensures that outliers, a common occurrence in social networks, do not significantly impact our results. (See the Appendix for more details.)

6 Results

6.1 RQ1: Legislators' Visibility by Party, Gender, Ethnicity, Posting Frequency, State

Table 4 shows the effect sizes for the Mann-Whitney U test. We only report the results for V^{I_P} here, the results for $25th$, $50th$, and $75th$ percentile are added in the Appendix along with 95% CI for Table 4. Overall, the visibility of legislators differs significantly based on party, gender, and posting frequency on both platforms and also for ethnicity on Twitter. Interestingly, Democrats and women appear to have higher visibility on Twitter but Republicans and men have higher visibility on FB ($p < 0.001$). White legislators also receive more attention on Twitter. On both platforms, legislators with higher posting activity have higher visibility. Figure 6 shows the mean V^{I_P} across US states for Twitter and FB. the visibility of legislators also differs based on their state representation. The most visible state on Twitter is New Mexico and Mississippi on FB. The posting rate of legislators is the second highest for New Mexico on Twitter which could be a possible explanation for the high visibility[26].

[26]For instance on FB, Mississippi Republican senator Chris McDaniel has a remarkably high engagement which dominates the vis-

IVs	Twitter	FB	Overlapping (OL) Twitter	FB
Party *(Rep vs. Dem)*	0.239***	-0.299***	0.189***	-0.347***
Gender *(Men vs. Women)*	0.076***	-0.089***	0.009	-0.128***
Ethnicity *(White vs. Non-White)*	-0.067**	-0.031	-0.023	-0.062
Posting Freq. *($\leq$ vs. $>$ median)*	0.457***	0.435***	0.368***	0.418***

$^{\cdot}p < 0.1, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table 4: RQ1 Effect sizes

We also look at the visibility of overlapping users (OL) on these platforms. Similar to prior results, Republicans and men receive higher engagement on FB and Democrats on Twitter. However, we do not find any significant difference across gender and ethnicity on Twitter for OL. Our results suggest that there exists cross-platform differences in how audiences engage with political content.

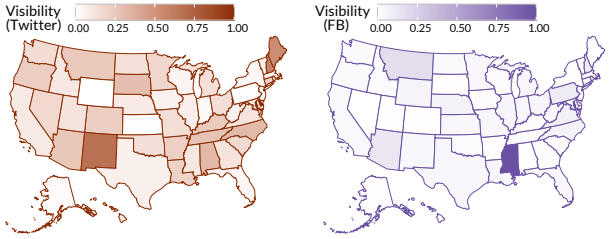

Figure 6: **Mean visibility of legislators from states.** The visibility (V^{I_P}) (normalized between 0-1) of legislators differ based on their state representation and platform.

6.2 RQ2: Factors Related to Visibility

In RQ2 we look at factors related to legislators' platform visibility. The results for all DVs are similar but we only report results for V^{I_P} in Table 5 (See Appendix for 95% CI). We are unable to estimate the interaction effect between party and low-credibility posts for FB due to insufficient low-credibility posts from Democrats. Party, gender, and post frequency are significantly correlated with visibility after controlling for other variables on both platforms. Republicans and men are more likely to garner higher visibility on FB whereas the opposite is true for Twitter. Higher posting activity is also related to higher visibility on both platforms. On Twitter, the number of followers and centrality in the intra-legislator network are not correlated with the legislators' visibility. Interestingly, there is a positive relation between legislators' network visibility and visibility.

As shown in Table 5, the volume of low-credibility posts is positively associated with legislators' visibility on FB

ibility term for the state.

IVs	Twitter	FB	Overlapping (OL) Twitter	FB
Party [Rep]	-0.133**	0.423***	-0.142*	0.512***
Gender [Men]	-0.078*	0.090**	-0.011	0.113**
Ethnicity [White]	0.020	0.015	0.020	0.040
#posts	0.401***	0.477***	0.409***	0.499***
#followers	-0.028	-	-0.028	-
Centrality	-0.020	-	-0.029	-
Network visibility	0.040*		0.027	-
#Low-Credible	-0.268**	0.079***	-0.335*	-0.037*
#Uncivil	0.148***	-	0.152***	-
Party [Rep] x #Low-Credible	0.299**	-	0.351**	-
R^2	0.289	0.382	0.291	0.386

$^{\cdot}p < 0.1, ^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

Table 5: RQ2 Regression Results

($p < 0.001$) but has an opposite effect ($p < 0.01$) on Twitter. However, visibility is positively correlated with the interaction between party and low-credibility posts on Twitter, i.e., one standard deviation increase in low-credibility posts by Republicans is associated with a 0.299 standard deviation increase in visibility . Thus, posting low-credibility content is related to higher visibility for only Republicans, otherwise it has a negative association on Twitter. Posting more uncivil content also increases the visibility of legislators on Twitter ($p < 0.001$). These results suggest that posting harmful content is associated with legislators' platform visibility.

We find similar results for OL, i.e., men and Republicans relate to higher visibility on FB, and Democrats on Twitter, again suggesting the cross-platform differences. Posting uncivil content on Twitter is positively associated with increased visibility, while posting low-credibility content is negatively associated with it. Moreover, visibility is positively associated with the interaction term between party and low-credibility posts. Interestingly however, posting low-credibility content is related to decreased visibility for the OL on FB similar to Twitter. Therefore, legislators who post content on both platforms have different communication strategies in comparison to the general legislator populations on those platforms which may be attributed to audience preferences across platforms.

Thus, posting harmful content is related to the visibility of the legislators. But the observed correlation may be a *spurious* one due to confounders, such as content in similar topics. Next, we analyze whether incivility or low-credibility of a post *impacts* its visibility.

6.3 RQ3: Observed Causal Impact of Incivility and Low-credibility on Visibility

Figure 7 shows the CATE estimates and 95% CI after matching. CATE is the expected change in the overperformance of a post if it contains low-credibility or incivility. For FB, we find that Republican legislators receive higher attention on posting low-credibility content. Similar results hold when we look at Extreme and OL Republicans. Interest-

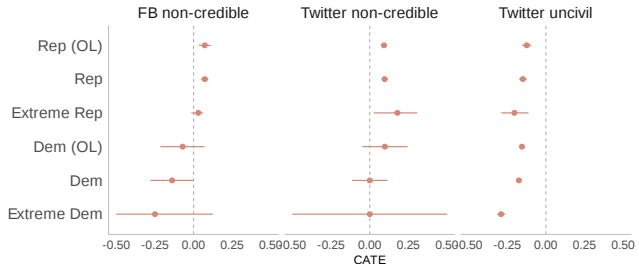

Figure 7: **Observed causal impact of low-credibility and incivility on legislators' content visibility.** After controlling for confounders, low-credibility has a positive impact on content visibility for Republicans on both platforms, but a negative effect for Democrats on FB. Incivility significantly hinders content visibility for all subgroups on Twitter.

ingly, the visibility of Democrats decreases when they post low-credibility content on FB. There are no effects for Extreme and OL Democrat subgroups.

For Twitter, similar to FB, Republicans, including their OL and Extreme subgroups receive higher visibility on posting low-credibility content. The effect size is higher for Extreme Republicans, i.e., the more conservative a legislator is the higher attention they receive on posting misinformation. For Democrats, however, we do not find any significant effects. The difference in outcomes for Democrats across Twitter and FB could either be due to their distinct behaviors (e.g., content) and/or audience preferences across these platforms. The average number of posts containing low-credibility content by Democratic legislators is higher on Twitter than on FB as shown in Table 3. This may suggest that Democrats post less low-credibility content on FB since their visibility is penalized and/or vice-versa.

The cross-platform analysis results suggest that there are asymmetries in how the Republican and Democratic party audiences engage with low-credibility content online and these asymmetries hold across platforms within and/or between parties. The former are more likely to engage with misinformation as shown by our results.

For uncivil content on Twitter, we find that both Democrats and Republicans, including their subgroups, receive lower engagement on posts containing uncivil language. The negative effects are higher for Democratic legislators compared to Republicans. The effects are also higher for Extreme legislators from both parties when compared to their party baselines. This implies that audiences irrespective of their partisan preferences engage less with uncivil content posted by legislators on Twitter and the legislators towards the extreme political spectra are penalized more.

7 Discussion

We analyzed the factors influencing the visibility dynamics of US state legislators by conducting a cross-platform analysis across Twitter and FB to understand different audience behaviors across these platforms. We showed that legislators' visibility varies based on their demographics, party, and posting frequency. Democrats have higher visibility on Twitter whereas Republicans have higher visibility on FB. Moreover, the regression analysis showed that a strong correlation exists between party and visibility, i.e., Democrats are associated with higher engagement on Twitter and Republicans on FB. These results also hold for the overlapping legislators, suggesting that audiences across these platforms engage with political content differently. Posting harmful content is also associated with legislators' visibility. Low-credibility content is related to increased visibility on FB, but decreased visibility on Twitter. Taking the effect of party into account, low-credibility posting correlates with higher visibility for Republicans. Uncivil posting is also correlated with higher account-level visibility on Twitter.

We further analyzed whether the increased visibility is due to posting harmful content after controlling for confounding factors such as demographics, party, topics, and time. Low-credibility garners more attention for Republicans on both platforms. However for Democrats, low-credibility reduces their content visibility on Twitter. These results highlight the partisan asymmetries in how low-credibility content receives attention online. Existing works have shown population asymmetries (Rao et al. 2022); our study further reveals attention disparity due to political actors' party affiliation. Higher online visibility provides a greater opportunity to influence public opinion (e.g., by gaining followers, impacting ranking algorithms). Politicians may post higher low-credibility content for political gains which may have implications for platform moderation.

Unlike low-credibility, incivility decreases the visibility of legislators' posts on Twitter for both Republicans and Democrats. The negative effects are more pronounced for Democrats compared to Republicans as well as for Extreme legislators, suggesting that audiences prefer to engage less with uncivil content irrespective of partisan preferences. Prior research has shown that uncivil content receives more negative reactions (Coe et al. 2014) owing to its emotionally charged language. The lower visibility may be attributed to the lack of expressing negative sentiments on Twitter, but further research is needed to confirm this. Our results highlight the cross-platform differences as well as asymmetries in how Democratic and Republican party audiences engage with political content online which is aligned with previous literature (Kelm 2020; Sahly et al. 2019) .

We show that harmful content is associated with legislators' online visibility. Such observed associations may be spurious, and other factors may contribute to their visibility. For instance, posting uncivil content has a positive association with visibility on Twitter but incivility has a negative impact on content visibility after controlling for confounders. Other factors may include the topics of their posts or simply the post timing. External factors (e.g., offline campaigns, media presence) can also contribute to their online visibility which is out of scope for this study. Moreover, there may be spillover effects from legislators' network visibility as hinted in our RQ2 results. Nevertheless, this study sheds light on some of the factors influencing legislators' overall as well as content visibility, but more research is needed to fully understand their online visibility dynamics.

Limitations and Future Work. Our study has certain limitations. We only look at the years 2020 and 2021—the generalizability to other periods remains uncertain. Our method of identifying low-credibility content was conservative which could have led to certain biases in our sampling. We demonstrated the feasibility of addressing the representation biases in Section 5; however, it is uncertain whether we were able to fully correct for them. Furthermore, we only identified uncivil and low-credibility posts based on the textual content but do not consider other forms of content (e.g., images) which may also contain harmful information.

We only looked at the visibility based on total interactions received on posts without discriminating between positive and negative reactions. Future work can study the impact of harmful content on positive and negative visibility separately to get a more nuanced understanding of public receptivity. It would also be interesting to look at the impact of posting harmful content on different types of engagement (e.g., Likes vs. Retweets). Another possible extension could be adapting our causal study design for continuous treatment (e.g., how visibility is affected by the degree of incivility).

Acknowledgements

The authors would like to acknowledge support from NSF #2318461, AFOSR, and Pitt Cyber Institute's PCAG awards. The research was partly supported by Pitt's CRC resources (RRID:SCR 022735 through NIH #S10OD028483). Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

References

Arhin, K.; et al. 2021. Ground-Truth, Whose Truth?–Examining the Challenges with Annotating Toxic Text Datasets. *arXiv preprint arXiv:2112.03529*.

Bahramirad, S. 2022. Virtual forums for public accountability: How internet and communication technologies are influencing citizen interactions with a local government. *CJAS*, 39(3): 313–327.

Bennett, W. L.; et al. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication*, 33(2).

Bode, L.; et al. 2016. Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing*, 15(4): 311–332.

Bodrunova, S. S.; and Blekanov, I. S. 2021. A self-critical public: Cumulation of opinion on Belarusian oppositional YouTube before the 2020 protests. *Social Media+ Society*, 7(4): 20563051211063464.

Bodrunova, S. S.; et al. 2021. Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube. *Media and Communication*, 9: 181–194.

Boulianne, S.; and Larsson, A. O. 2023. Engagement with candidate posts on Twitter, Instagram, and Facebook during the 2019 election. *New Media & Society*, 25(1): 119–140.

Bovet, A.; et al. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1): 7.

Coe, K.; et al. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4): 658–679.

Crowdtangle. 2024. How do you calculate overperforming scores? https://help.crowdtangle.com/en/articles/2013937-how-do-you-calculate-overperforming-scores. Accessed: 2023-12-11.

Cuan-Baltazar, J. Y.; et al. 2020. Misinformation of COVID-19 on the internet: infodemiology study. *JMIR public health and surveillance*, 6(2): e18444.

Eberl, J.-M.; et al. 2020. What's in a post? How sentiment and issue salience affect users' emotional reactions on Facebook. *Journal of Information Technology & Politics*, 17(1).

Enli, G. S.; and Skogerbø, E. 2013. Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, communication & society*, 16(5): 757–774.

Feder, A.; et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the ACL*, 10: 1138–1158.

Ferrara, E.; et al. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday*.

Friggeri, A.; et al. 2014. Rumor cascades. In *ICWSM*, volume 8, 101–110.

Frimer, J. A.; et al. 2023. Incivility is rising among American politicians on Twitter. *SPPS*, 14(2): 259–269.

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

Goovaerts, I.; et al. 2020. Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication*, 37(6).

Grant, W. J.; et al. 2010. Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian journal of political science*, 45(4): 579–604.

Halpern, D.; et al. 2019. From belief in conspiracy theories to trust in others: Which factors influence exposure, believing and sharing fake news. In *SCSM 2019*. Springer.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify. Accessed: 2025-04-10.

Hasan, R.; et al. 2022. The Impact of Viral Posts on Visibility and Behavior of Professionals: A Longitudinal Study of Scientists on Twitter. In *ICWSM*, volume 16, 323–334.

Hua, Y.; et al. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *CHI 2020*, 1–13.

Johansson, F.; et al. 2016. Learning representations for counterfactual inference. In *ICML*, 3020–3029. PMLR.

Johnson, J. 2018. The self-radicalization of white men:"Fake news" and the affective networking of paranoia. *Communication Culture & Critique*, 11(1): 100–115.

Jungherr, A. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1): 72–91.

Kelm, O. 2020. Why do politicians use Facebook and Twit-

ter the way they do? The influence of perceived audience expectations. *SCM Studies in Communication and Media*, 9(1): 8–34.
Kim, J. W.; et al. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6).
Kim, T.; et al. 2022. Attention to the COVID-19 Pandemic on Twitter: Partisan Differences Among US State Legislators. *Legislative studies quarterly*, 47(4): 1023–1041.
Koch, B.; et al. 2021. Deep Learning for Causal Inference.
Kreiss, D. 2016. Seizing the moment: The presidential campaigns' use of Twitter during the 2012 electoral cycle. *New media & society*, 18(8): 1473–1490.
Kyriakidou, M.; et al. 2021. Journalistic responses to misinformation. *The Routledge Companion to Media Disinformation and Populism*, 529–537.
Lasser, J.; et al. 2022. Social media sharing of low-quality news sources by political elites. *PNAS nexus*, 1(4): pgac186.
Liu, Y.; et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
Mihailidis, P.; et al. 2021. The cost of disbelief: Fracturing news ecosystems in an age of rampant media cynicism. *ABS*, 65(4): 616–631.
Mutz, D. C. 2007. Effects of "in-your-face" television discourse on perceptions of a legitimate opposition. *APSR*, 101(4): 621–635.
Pelletier, M. J.; et al. 2021. Fexit: The effect of political and promotional communication from friends and family on Facebook exiting intentions. *Journal of Business Research*, 122: 321–334.
Pennycook, G.; and Rand, D. G. 2021. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The HKS Misinformation Review*.
Pryzant, R.; et al. 2020. Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919*.
Rao, A.; et al. 2022. Partisan asymmetries in exposure to misinformation. *Scientific Reports*, 12(1): 15671.
Roberts, M. E.; et al. 2020. Adjusting for confounding with text matching. *AJPS*, 64(4): 887–903.
Sahly, A.; et al. 2019. Social media for political campaigns: An examination of Trump's and Clinton's frame building and its effect on audience engagement. *Social Media+ Society*, 5(2): 2056305119855141.
Serrano-Puche, J. 2021. Digital disinformation and emotions: exploring the social risks of affective polarization. *International Review of Sociology*, 31(2): 231–245.
Shi, C.; et al. 2019. Adapting neural networks for the estimation of treatment effects. *NeurIPS*, 32.
Shmargad, Y.; et al. 2022. Social norms and the dynamics of online incivility. *Social Science Computer Review*, 40(3).
Shor, B.; and McCarty, N. 2011. The ideological mapping of American legislatures. *APSR*, 105(3): 530–551.
Squire, P.; et al. 2019. *State legislatures today: Politics under the domes*. Rowman & Littlefield.
Sridhar, D.; and Getoor, L. 2019. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*.
Stier, S.; et al. 2020. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. In *Studying Politics Across Media*, 50–74. Routledge.
Stryker, R.; et al. 2016. What is political incivility? *Communication monographs*, 83(4): 535–556.
Tai, Y. C.; et al. 2023. Official yet questionable: examining misinformation in US state legislators' tweets. *Journal of Information Technology & Politics*, 1–13.
Toraman, C.; et al. 2022. BlackLivesMatter 2020: An analysis of deleted and suspended users in Twitter. In *WebSci 2022*, 290–295.
Twitter-team. 2024. the-algorithm. https://github.com/twitter/the-algorithm?tab=readme-ov-file. Accessed: 2025-04-10.
Veitch, V.; et al. 2020. Adapting text embeddings for causal inference. In *UAI*, 919–928. PMLR.
Vosoughi, S.; et al. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.
Yang, Y.; et al. 2023. Visual misinformation on Facebook. *Journal of Communication*, jqac051.
Zollo, F.; et al. 2015. Emotional dynamics in the age of misinformation. *PloS one*, 10(9): e0138740.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. See Section 6.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. See Section 4 and 5.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. See Section 3.
   (e) Did you describe the limitations of your work? Yes. See Section 7.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. See Section 8.
   (g) Did you discuss any potential misuse of your work? Yes. We discuss potential mis-intepretation of our conclusions in Section 8.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We will release the dataset in the future.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes. See Section 2.
   (b) Have you provided justifications for all theoretical results? Yes. See Sections 6 and 7.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes. See Section 2.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes. We conduct robustness check to consolidate our findings. See Appendix
   (e) Did you address potential biases or limitations in your theoretical framework? Yes. See Sections 7 and 8.
   (f) Have you related your theoretical results to the existing literature in social science? Yes. We discuss our theoretical foundations in Section 3.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes. We discuss the implications in Section 6 and 7.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A.
   (b) Did you include complete proofs of all theoretical results? N/A.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We will release them later for the anonymous review.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. See Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. We reported error bars whenever possible.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. See Section 5.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. See Section 8.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes. See Section 3.
   (b) Did you mention the license of the assets? N/A.
   (c) Did you include any new assets in the supplemental material or as a URL? No.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. See Section 8.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. We conform to the social media policies. See Section 8.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? No. We need to conform to social media platform policies sharing our curated data. That means, we can only share the public post IDs, without the data content itself.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? We will do our best releasing the data without breaching the social media platforms' policies.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? N/A.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A.

(d) Did you discuss how data is stored, shared, and deidentified? N/A.

Ethics Statement

Data. We collect data from two sources, Twitter and FB. For Twitter, the data was collected using Twitter's Official API v2.0 before rate limitations were imposed (i.e., March 2023). For FB, we collect data using CrowdTangle's official API by following their terms of service. All the data are publicly posted and available for viewing without restrictions.

We ensure that the Dragonnet model can effectively deconfound the covariate representation space for treated and non-treated samples based on our qualitative analysis and performance metrics. The classification errors from Dragonnet model are less likely to affect our results since we do not use the model predictions to calculate CATE. However, misclassification may still impact the deconfounding quality, resulting in confounding from textual content that we cannot measure, unlike other covariates. Our study suggests that public figures sharing harmful content on social media has significant consequences. We showed that when low-credibility content is posted by public figures, the combination of user behavior (interacting with the posts) and platform mechanisms (e.g., feed recommendation algorithms) can result in increased visibility for such content. Our findings should not be interpreted as an encouragement to spread such low-credibility content; instead, they should serve as a warning that there may be incentives for political or elite actors to do so. Moreover, our study has focused on the behavior of US subnational politicians on two specific platforms—Twitter and FB. The results may not be generalizable to other platforms and social media users including the activities of political opinion leaders and media elites from other countries, or even US national politicians, due to several factors—media scrutiny, platform moderation rules, and public perception to name some. More research is needed to understand whether our results generalize to other settings.

Appendix

Data Collection: For Twitter data, we employed a comprehensive approach, drawing from various sources, including existing datasets (Kim et al. 2022), and conducting searches on Google, Twitter, Wikipedia, legislators' official websites, campaign sites, and Ballotpedia, to compile the accounts and demographic information of state legislators. This meticulous strategy facilitated the manual identification and verification of legislators' Twitter accounts. Subsequently, we refined our dataset to only include legislators who served between 2020 and 2021, determined by their tenure in office. It is important to acknowledge that the completeness of our data was affected by factors such as inactive or inaccessible accounts after legislators left office.

Our initial approach to collecting FB data involved gathering posts from accounts bearing names indicative of belonging to state legislators. Subsequent verification against information from Ballotpedia allowed us to filter out non-legislator accounts. To address data gaps, we conducted three successive rounds of data recollection in April 2022, March 2023, and April 2023. The successive rounds allowed us to capture posts from accounts previously overlooked. However, similar to Twitter data, numerous accounts had become inaccessible, largely due to campaign or official accounts no longer being active.

Despite these efforts, we encountered a significant challenge with FB data collection concerning accounts that were not listed on Ballotpedia. Although we attempted to identify missing accounts using keyword searches, achieving a perfect match with the legislator information on Ballotpedia was difficult. This limitation resulted in a higher rate of mismatch in the mapped attributes of FB accounts, i.e., around 70.2% of legislators could be mapped to their attributes for FB. Table 7 shows the statistics of our FB dataset after mapping the legislators to their attributes. The trends are similar to that in Table 2 which suggests that no or minimal biases were introduced during our mapping process.

Figure 8 shows the breakdown of ethnicity and gender by party and for OL.

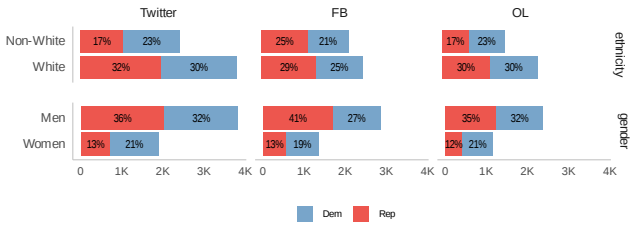

Figure 8: Ethnicity and gender by party, platform, and for overlapping users. Our dataset has higher representation of men and White legislators on both platforms.

Assessing Post's Civility. For a study like ours it is hard to interpret the continuous toxicity scores returned by the Detoxify model. So we follow the common practice in literature to convert the toxicity scores to binary based on a threshold (Hua et al. 2020). It is important to have a threshold for the toxicity for our particular dataset because uncivil language may evolve over different topics and time. To estimate an ideal cutoff for the toxicity score, we manually annotate a sample of 300 posts as uncivil or civil using stratified sampling such that more samples are included at higher toxicity scores. Three annotators labeled all 300 samples based on the aforementioned definition of toxic language. Since Cohen's Kappa suffers from imbalanced label distributions, we compute the pairwise agreement scores for the percentage of total samples agreed upon by the annotators, which ranges between 66.3-85.3%. The final labels are decided according to the majority vote. Based on the ROC (AUC $= 0.81$), we choose the cutoff for toxicity as 0.82, i.e., posts having a score above 0.82 are considered uncivil. This is similar to previous works using Detoxify or Perspective API which have a threshold between 0.5-0.9 (Hua et al. 2020), though we acknowledge that our threshold is more towards the conservative side which is done to ensure that uncivil posts are selected with high precision. Table 6 shows examples of uncivil posts by legislators on Twitter and FB.

The number of uncivil posts on FB[27] is much lesser compared to Twitter which shows that the political communication styles are different on two platforms. This finding also resonates with previous studies which have suggested that FB is used more for broadcasting purposes whereas Twitter is used more for direct communication (Enli and Skogerbø 2013). Based on this, it is reasonable that there are very less uncivil posts on FB since the language used on FB tends to be more formal. We further verify this by measuring the readability scores of author's posts on these two platforms using Flesch–Kincaid grade level. The median readability of legislators is 11.04 (i.e., the text is written at a level suitable for someone who has completed the 11th grade in the US education system) on FB and 9.55 on Twitter. This shows that the language used on FB is indeed more formal and potentially the reason why it is more civil.

Measuring Visibility. The visibility metric is designed to capture the overall engagement on the platforms and hence our metric includes all/most of the elements used to calculate engagement on the respective platforms. We further analyze the contribution of individual visibility elements on each platform. On Twitter, Likes contribute 2.3% and Retweets contribute 97.7% to the overall interactions. Reply and quote consist of a small percentage of the interactions. On FB however, Likes contribute 51.8%, Shares contribute 17.0% and Comments contribute 13.2% to the overall interactions. This suggests that audiences engage differently with content on Twitter and FB, e.g., retweeting is most dominant form of engagement on Twitter whereas Liking for FB. Moreover, the interpretation of individual elements may also be different across platforms, for instance, audiences may engage with Like on Twitter differently than Like on FB simply due to different icons, therefore a direct comparison of individual engagement metrics may not be justified. So, by only including individual elements, the visibility metric may not be able to capture the level of engagement on these plat-

[27]The number of uncivil posts is still low at other cutoffs, for e.g., cutoff = 0.1 yields around 2,690 uncivil posts

platform	party	text
FB	Rep	*While millions of Americans have yet to receive their stimulus checks, a new study reveals that $4.38 billion of the new round will go right into the pockets of illegal immigrants. Another dumb idea and stupid stupid policy. What is wrong with these people?*
FB	Dem	*"I didn't think it would be this ridiculous. It's embarrassing to be a state senator at this point," Paul Boyer said of partisan recount. Yes it does make you look like idiots. Wasting time & resources on #TrumpsBigLie*
Twitter	Rep	*We are at the start of one of the LARGEST recessions in American history, which will DESTROY many lives, and people are still in favor of partial lockdownshow stupid could you possibly be! Bunch of damn fools.*
Twitter	Dem	*You are a blithering idiot. Who gives a shit about the VP. Vote for Trump and thousands upon thousands will die.*

Table 6: Examples of Uncivil Posts on Twitter and FB, by party

	FB		
party	**#users**	**#tweets**	**Int**
Dem	1588	171K	61.0
Rep	1718	152K	114.0

Table 7: Descriptive statistics for FB dataset after mapping, showing the number of legislators, posts, and median interactions received per post by party.

			OL	
IVs	**Twitter [CI]**	**FB [CI]**	**Twitter [CI]**	**FB [CI]**
Party	0.239	-0.299	0.189	-0.347
	[0.202, 0.277]	[-0.337, -0.259]	[0.141, 0.240]	[-0.397, -0.298]
Gender	0.076	-0.089	0.009	-0.128
	[0.033, 0.117]	[-0.132, -0.046]	[-0.044, 0.059]	[-0.182, -0.072]
Ethnicity	-0.067	-0.031	-0.023	-0.062
	[-0.108, -0.027]	[-0.079, 0.015]	[-0.073, 0.029]	[-0.128, 0.001]
Posting	0.457	0.435	0.368	0.418
Freq.	[0.427, 0.492]	[0.398, 0.470]	[0.323, 0.414]	[0.370, 0.461]

Table 8: 95% CI for Table 4

forms, making the interpretation of the metric harder for a cross-platform study.

We further examine the possibility of the visibility metric being dominated by a single element by testing if the underlying distributions are similar for the total interactions and the most dominant element. According to our KS tests, for both Likes on FB and Retweets on Twitter, we find that the distributions are significantly different (p-value$<$0.05) compared to the total interactions, suggesting that other elements also contribute to the overall engagement on both platforms and hence need to be included in the visibility measure.

DV Transformation for RQ2. To satisfy assumptions of linear regression in RQ2, we transform all variables to be close to normal distributions using the "bestNormalize" R package. For Twitter, we transform variables as follows: V^{I_P} (Yeo-Johnson), #posts (Box Cox), #Misinfo (sqrt), #Uncivil (sqrt), Network visibility (Center+scale), #followers (Box Cox), Centrality (None). For FB, we transform variables as follows: V^{I_P} (Yeo-Johnson), #posts (Box Cox), #Misinfo (sqrt).

IVs	25^{th}	50^{th}	75^{th}
Party	0.120^{***}	0.160^{***}	0.160^{***}
Gender	0.016	$0.036_{.}$	0.060^{**}
Ethnicity	0.004	-0.027	-0.054^{**}
Posting Freq.	0.255^{***}	0.354^{***}	0.386^{***}

Table 9: Robustness analysis of RQ1 results for $25th$, $50th$, and $75th$ percentile visibility on FB

IVs	25^{th}	50^{th}	75^{th}
Party	-0.293^{***}	-0.300^{***}	-0.291^{***}
Gender	-0.078^{***}	-0.088^{***}	-0.088^{***}
Ethnicity	-0.048	$-0.047_{.}$	-0.032
Posting Freq.	0.376^{***}	0.402^{***}	0.424^{***}

Table 10: Robustness analysis of RQ1 results for $25th$, $50th$, and $75th$ percentile visibility on Twitter

RQ1 Tables. Table 8 shows the 95% CI for Table 4. Tables 10 and 9 show the results for $25th$, $50th$, and $75th$ percentile visibility for Twitter and FB respectively.

RQ2 Tables. Table 12 shows the 95% CI for Table 5.

Benchmarks for $thres$, w. Figure 9 shows the ECDF plots for daily mean interactions received by legislators on Twitter and FB, by party. For Twitter and FB, the medians are close to 10 and 100 respectively. So we select $thres = 10$ for Twitter and $thres = 100$ for FB. We do not have different $thres$ across parties since medians are similar across parties on both platforms.

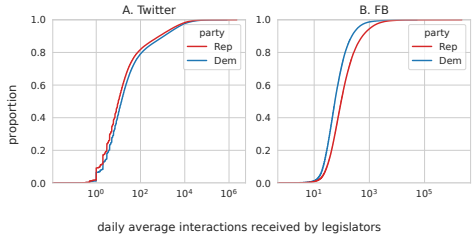

Figure 9: ECDF plots for daily mean interactions received by legislators on Twitter and FB, by party.

Figure 10 shows the ECDF plots for daily number of posts by legislators on Twitter and FB, by party. Again the median posting rates are similar for Republicans and Democrats on

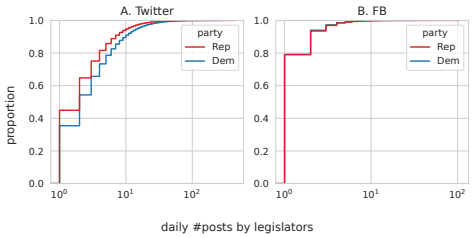

Figure 10: ECDF plots for daily number of posts by legislators on Twitter and FB, by party.

	AUC		Macro F1	
	overall	Extreme	overall	Extreme
Twitter uncivil	0.74	0.72	0.69	0.66
Twitter non-credible	0.73	0.74	0.66	0.68
FB non-credible	0.80	0.88	0.74	0.82

Table 11: Dragonnet performance

both Twitter and FB. The median daily post count on Twitter is 2 and 1 on FB. To ensure that we have a sufficient number of posts per legislator to compute the overperforming scores and simultaneously account for temporal variation in our data, we select $w = 14$ for both Twitter and FB.

Dragonnet Model Description. Dragonnet uses feed-forward neural networks to learn balanced representations of the data such that each head models a separate potential outcome, a third propensity head predicts the propensity (π) of being treated, and a free nudge parameter ϵ. The π and ϵ parameters are used to re-weight the outcomes to provide lower biased estimates of the *Conditional Average Treatment Effect* (CATE). The error gradients from the two outcome modeling heads are propagated back to the shared representation layers of the Dragonnet model to learn the covariate representation, i.e., $\phi(X)$. The representation layers learn a balanced representation of the data since the model objective is to predict both outcomes and each outcome modeling head learns a function of the transformed covariate representation, i.e., $Y(T) = h(\phi(x), T)$. The CATE from Dragonnet predictions is estimated as follows,

$$CATE = \frac{1}{N}\sum_{i}^{N}(Y_i^*(1) - Y_i^*(0)) \quad (4)$$

where,

$$Y_i^* = \hat{Y}_i + (\frac{T_i}{\pi(\phi(X_i), 1)} - \frac{1 - T_i}{\pi(\phi(X_i), 0)}) \times \epsilon \quad (5)$$

where $\hat{Y}(1)$ and $\hat{Y}(0)$ are predictions returned by the two outcome modeling heads respectively, π is the predicted propensity of a sample being treated, and sample size N.

Dragonnet Model Performance. Figure 11 shows the ROC curves for Twitter incivility, Twitter low-credibility and FB low-credibility Dragonnet models. The AUC and Macro F1-scores[28] are reported in Table 11.

[28]F1-scores reported at optimal cutoff

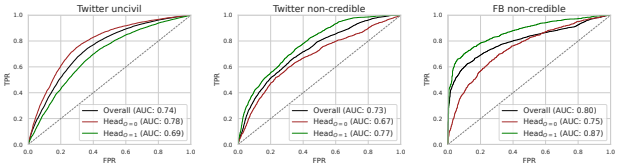

Figure 11: ROC curves showing Dragonnet performance of overall as well as two outcome modeling heads, for Twitter incivility, Twitter and FB low-credibility respectively.

			OL	
IVs	**Twitter**	**FB**	**Twitter**	**FB**
Party [Rep]	-0.133	0.423	-0.142	0.512
	[-0.219, -0.046]	[0.359, 0.487]	[-0.252, -0.029]	[0.431, 0.594]
Gender [Men]	-0.078	0.090	-0.011	0.113
	[-0.149, -0.007]	[0.029, 0.152]	[-0.101, 0.079]	[0.033, 0.194]
Ethnicity [White]	0.020	0.015	0.020	0.040
	[-0.061, 0.100]	[-0.056, 0.086]	[-0.079, 0.118]	[-0.054, 0.133]
#posts	0.401	0.477	0.409	0.499
	[0.345, 0.455]	[0.443, 0.510]	[0.337, 0.481]	[0.459, 0.538]
#followers	-0.028	-	-0.028	-
	[-0.076, 0.021]		[-0.091, 0.034]	
Centrality	-0.020	-	-0.029	-
	[-0.060, 0.019]		[-0.079, 0.021]	
Network visibility	0.040		0.027	-
	[0.002, 0.080]		[-0.020, 0.076]	
#Low-Credible	-0.268	0.079	-0.335	-0.037
	[-0.466, -0.070]	[0.047, 0.110]	[-0.601, -0.069]	[-0.074, 0.000]
#Uncivil	0.148	-	0.152	-
	[0.101, 0.196]		[0.092, 0.213]	
Party [Rep] x	0.299	-	0.351	-
#Low-Credible	[0.103, 0.496]		[0.087, 0.614]	
R^2	0.289	0.382	0.291	0.386

Table 12: 95% CI for Table 5

Covariate Balance for Matching. We employ 1:1 matching such that each treated sample is matched to one untreated sample. We use Nearest Neighbour matching based on Euclidean distance between the deconfounded tweet embeddings. We find matches for 9677 (64.0%) uncivil tweets, 3957 (73.5%) low-credibility tweets, and 1583 (61.1%) low-credibility FB posts using a distance cutoff of 0.1 to maximize the covariate overlap. This gives us balanced representations of the observed covariates across the treated and untreated samples as shown in Figure 6. We compute the standardized differences for each of the covariates before and after matching. A score between -0.1-0.1 indicates balance.

Effect of Outliers Outliers are common in social networks, but their impact on analysis results and conclusions can vary. In our dataset, outliers may exist in terms of posting volume and engagement received due to the scale-free distributions (refer to Fig 12). However, we have taken measures to ensure these outliers do not significantly impact our results.

For RQ1, we used a non-parametric statistical test, the Mann-Whitney U test, to compare distributions, which is

robust to outliers. For RQ2, we used non-linear transformations on study variables to minimize the impact of outliers in the regression analysis.

In RQ3, we matched accounts with similar characteristics in the de-confounded embedding space, such as similar posting rates. This process either discarded outliers if no adequate match was found or retained them if an adequate match was identified. This matching step ensured the balance of the covariates before running the estimation of the Conditional Average Treatment Effect (CATE), further reducing the impact of outliers.

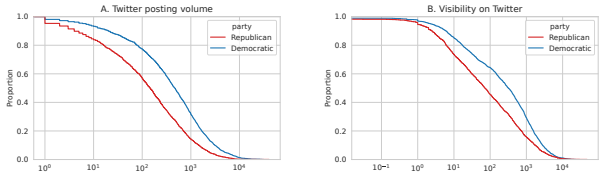

Figure 12: Scale-free distributions for (A) Posting volumes and (B) Visibility of legislators on Twitter. The distributions are also similar for FB.