Twits, Toxic Tweets, and Tribal Tendencies: Trends in Politically Polarized Posts on Twitter

HANS W. A. HANLEY, Stanford University, USA
ZAKIR DURUMERIC, Stanford University, USA

Social media platforms are often blamed for exacerbating political polarization and worsening public dialogue. Many claim that hyperpartisan users post pernicious content slanted toward their political views, inciting contentious and toxic conversations. However, what factors are actually associated with increased online toxicity and negative interactions? In this work, we explore the role that partisanship and affective polarization play in contributing to toxicity both at the individual user level and at the topic level on Twitter/X. To do this, we train and open-source a DeBERTa-based toxicity detector that outperforms the Google Jigsaw Perspective API toxicity detector on the Civil Comments test dataset. After collecting 89.6 million tweets from 43,151 US-based Twitter/X users, we then examine how several account-level characteristics—including partisanship along the US left–right political spectrum—predict how often users post toxic content. Using a Generalized Additive Model (GAM), we find that both the diversity of views and the toxicity of other accounts with which users engage have a marked effect on users' own toxicity. Specifically, toxicity is correlated with users who engage with a wider array of political views. Performing topic analysis on the toxic content posted by these accounts using the large language model MPNet and a version of the DP-Means clustering algorithm, we find similar patterns across 5,288 topics, with users becoming more toxic as they engage with a broader diversity of politically charged topics.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Empirical studies in collaborative and social computing**; • **Information systems** → **World Wide Web**.

Additional Key Words and Phrases: Toxicity, Affective Polarization, Twitter, Online Communities

ACM Reference Format:
Hans W. A. Hanley and Zakir Durumeric . 2025. Twits, Toxic Tweets, and Tribal Tendencies: Trends in Politically Polarized Posts on Twitter. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW503 (November 2025), 40 pages. https://doi.org/10.1145/3757684

1 Introduction

Content Warning: This paper studies online toxicity. When necessary for clarity, this paper quotes user content that can be considered profane, politically inflammatory, and hateful.

Over the past decade, political polarization within the United States has increased substantially [14, 22, 43, 44, 46, 69]. Many people attribute the increase in division to social media, arguing that social media creates toxic political echo chambers where users become more politically polarized [138, 152]. Indeed, in several documented cases, political polarization and associated toxicity have negatively impacted platforms, online communities, and users, sometimes leading to users leaving platforms altogether [36]. While many studies have investigated the role that toxicity and political

Authors' Contact Information: Hans W. A. Hanley, hhanley@cs.stanford.edu, Stanford University, Stanford, California, USA; Zakir Durumeric, zakir@cs.stanford.edu, Stanford University, Stanford, California, USA.

ACM 2573-0142/2025/11-ARTCSCW503
https://doi.org/10.1145/3757684

polarization have had on the health of online communities [49, 110, 130, 143–145], there has been little work that investigates the role of toxicity, partisanship, and affective polarization (*i.e.*, the tendency to be negative to those with different political views and positive to those with similar political views) between individuals and at the topic level, the common means by which conversations take place on Twitter/X across multiple Twitter threads and on similar platforms like Bluesky and Threads [6, 45, 117, 151].

As argued in prior work, social media platforms tend to promote ideological congruence [54, 109, 131, 148] amongst their users and the content that they view, creating ideological echo chambers [136]. However, recent research has evidenced that this political alignment may lead to reduced toxicity on Facebook and Reddit [8, 54, 58, 108]. As found by An et al. [4], even within communities meant to encourage cross-political discussion on Reddit, users tend only to comment and interact with users of similar political beliefs as themselves. As argued in the social theories of Tajfel and Turner [139] and in Le et al.'s work [92], individuals categorize themselves and others into in-groups and out-groups, leading to these behaviors. Further, as found elsewhere, these interactions across political differences may indeed promote increased polarization and hostility [8, 106]. Most of this research, however, has largely focused on individual threads within stable communities on Reddit or pages on Facebook or across singular topics [39, 45, 156]. As argued by Prochaska et al. [113], conversations on platforms like Twitter mobilize geographically and politically diverse communities, allowing for the study of how politically heterogeneous interactions and platform-wide political discussions occur. This is of particular interest to the CSCW researchers given the current incomplete understanding of the rise of political polarization in digitally mediated spaces and its mirroring in the US as a whole [8, 113]. In this work, taking advantage of recent advancements in large language models [134], we thus seek to understand how these dynamics play out on an individual account level and across platform-wide political conversations. This enables us to study the dynamics of affective polarization within larger platform-wide conversations and to further understand how this polarization contributes to toxicity online. Concretely, to fully understand the intertwined relationship between toxicity, partisanship, and polarization, at the user and topic level in this work, we investigate:

(1) *RQ1: What are the relationships between partisanship, political polarization, and the tendency for politically engaged users to post toxic content?*
(2) *RQ2: How do the characteristics of users, including their partisanship, predict the toxicity of topics on Twitter/X?*

To answer these questions, we collect 89.6M tweets from 43.15K accounts throughout 2022. From these tweets, we measure the number of toxic tweets and the average toxicity of posts per user by designing and deploying our own DeBERTa-based [66] toxicity detection model. This newly designed model outperforms Google Jigsaw Perspective API [78], the gold-standard out-of-the-box classifier for identifying uncivil and toxic language (*e.g.*, insults, sexual harassment, and threats of violence [142]). Then, calculating each user's approximate political orientation using Correspondence Analysis [11] and performing fine-grained topic analysis using a large language model, we subsequently determine the interconnection between toxicity and political polarization at a user and topic level.

RQ1: User level Factors of Toxicity and the Role of Political Polarization. We first determine, using a Generalized Additive Model (GAM), some of the most significant features that predict the toxicity of content posted by individual Twitter accounts. We find that the most important feature that predicts an individual account's toxicity is the toxicity of the other accounts with which the user interacts. Namely, as users interact with other users who regularly tweet in a toxic manner, they themselves are more likely to tweet toxic content. We further find that while the position that

a user falls on the political spectrum *does not* have much bearing on the toxicity of their messages, the more that a given user interacts with users of different political orientations, the more likely their posts are to be toxic.

RQ2: Toxicity and Political Polarization in Toxic Topics. Having observed that users who interact with users of differing political views are more likely to be toxic, we examine this dynamic at a topic level. After identifying 5.5M English-language toxic tweets, we perform topic analysis using a fine-tuned version of the large language model MPNet and the DP-Means clustering algorithm [60]. Examining these topic clusters, we find that, in aggregate, the political orientation of users tweeting about a topic does not have a large effect on the topic's overall toxicity; rather, we find the political orientation of the users tweeting toxically about particular topics varies widely. Examining factors that predict each topic cluster's overall toxicity, we find, as largely expected, that high-toxicity topics often involve high-toxicity users. We further find that as individuals participate in a wider range of political topics, the toxicity of their tweets increases. Namely, we again identify at the topic level (as on a user level), a strong tribal tendency/affective polarization, with accounts acting negatively toward accounts of differing views.

Altogether, our work illustrates that, across a diverse set of users and topics, as engagement with toxic content and with a wider range of political views increases, so does average toxicity. In addition to open-sourcing a new toxicity classifier that achieves better accuracy than the Perspective API and several state-of-the-art decoder large language models [47, 73, 141], on the Civil Comments dataset, our work — one of the first to perform this analysis on a large-scale dataset of politically engaged users and across multi-thread topics not directly chosen by specific hashtags — illustrates how political polarization can negatively affect online communities and lead to increased divisiveness, regardless of the topic. We hope that this work helps inform future research into the role of polarization and toxic content in negatively affecting the health of online communities and intra-platform user interactions.

2 Background & Related Work

In this section, we detail several key definitions utilized within our study, provide background on Twitter, and finally present an overview of existing works that inform our study.

2.1 Terminology

We first provide some preliminary definitions of terms that form the basis of this work:

Online Toxicity and Incivility: We utilize the Perspective API's definition of online toxicity and incivility — "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion*" — given its extensive use in past studies of online toxicity [70, 88, 130, 155].

Political Partisanship: As in Barberá et al. [10] and other works [128, 129], we define US political partisanship along a unidimensional axis ranging from left-leaning (*i.e.*, liberal) to right-leaning (*i.e.*, conservative). While this limits our analysis, given the variety of political views within the US, as found by Poole and Rosenthal, most of the variation in US political ideology *is* along a unidimensional axis [112], and this assumption is fairly common in the literature.

Affective Polarization: Affective polarization is the tendency of individuals to distrust and be negative towards those of different political beliefs while being positive towards people of similar political views [34].

2.2 Twitter/X

Twitter/X is a microblogging website where users can post messages known as tweets — messages with at most 280 characters. Tweets themselves, while often just text, can also include hyperlinks,

videos, and other types of media [79]. Unless made private, tweets are publicly displayed on the Twitter platform, allowing anyone to see or reply to the message [81]. As of late 2022 (the time of this study), Twitter had approximately 238 million active daily users [30]. Many Twitter users get their daily news from the Twitter platform [3, 15, 140]. Despite the ability of anyone to gain and maintain a following on Twitter, several studies have found that political conversations are often dominated and guided by legacy media elites and celebrities [29]. We note that Twitter changed its name to X in mid-2023 [75], but for simplicity, we still refer to the platform as Twitter throughout this work.

2.3 Political Partisanship and Polarization Online

Various works have explored the role that individual users' political orientations play in interactions online. People, on the Internet and in their everyday interactions, tend to associate with like-minded individuals and Twitter is no exception [9, 11, 56, 71, 80, 115]. Several works have found that social media exacerbates this human tendency by creating political echo-chambers [136], where users' biases are reconfirmed and reinforced [5, 13, 25, 27]. Sunstein, Garrett et al., and Quattrociocchi et al. all argue that the "individualized" experience offered by social media companies comes with the risk of creating "information cocoons" and "echo chambers" that accelerate polarization [42, 116, 138]. Wojcieszak et al. [153], for example, determine that the majority of political discussions online are between participants who share the same viewpoint. Indeed, while the vast majority of Twitter users do not engage in political discussions, those who do are often highly politically polarized [152].

As found by Munson et al. [104], while some individuals seek views that are vastly different than their own, many also largely seek only affirming beliefs. Rogowski et al. [125] show that high ideological differences between individuals can lead to increased affective polarization; namely, if individuals are exposed to others with widely different beliefs, they increase their tendency to be negative toward those individuals and positive toward those who share their beliefs. Even more so, several recent research papers have found that social media can increase this rate of affective polarization [85, 137]. Cho et al. [24] find that exposure to social media content that attacks political figures can increase affective polarization. Most similar to our work, Bail et al. [8] show that exposure to different political beliefs online can increase polarization, particularly for right-leaning individuals.

In addition to polarization being amplified by social media, other works have found that this increased polarization can increase the spread of misinformation and toxic behavior [5]. Rains et al. [118], for instance, find that high polarization is a major factor in engendering online incivility and toxicity. Imhoff et al. [74], find that political polarization, on both sides of the political spectrum, is associated with beliefs in conspiracy theories.

2.4 Online Toxicity

Online toxicity (*e.g.*, doxing, cyberstalking, coordinated bullying, and political incivility) plagues social media platforms [20, 28, 89, 107, 142, 154]. As outlined by Thomas et al. [142], online toxicity is just one type of hate and harassment, which intersects with other negative online behaviors like misinformation and extremism. Brubaker et al. [17] find that trolls and bullies online are often motivated by a type of schadenfreude in spewing vitriol at other users. Similarly, Thomas et al. [142] find that abusers are often also motivated by political ideology, disaffection, and control [142]. For example, Flores-Saviaga [38] studied how users in the r/The_Donald were motivated to troll and abuse other Reddit users in support of then-Republican candidate Donald Trump in 2016. In addition to harming the target, online toxicity often has many negative downstream effects. Kim et al., Kwon et al., and Shen et al., find, for example, that online toxicity is a self-reinforcing behavior, with negative conversations increasing observers' tendency to also engage in incivility [82, 90, 132].

Other works have found that marginalized groups often receive disproportionate amounts of toxicity online [23, 121, 142]. Pew Research, for instance, found that Black adults reported higher incidences of name-calling, while women were more likely to experience sexual harassment. While toxicity can take many forms, in this work, we largely focus on toxic comments on Twitter.

2.5 Detecting Online Toxicity

Several works measure online toxicity using the Google Jigsaw Perspective API [78]. Saveski et al. [130], for example, utilize the Perspective API and find that many of the idiosyncrasies of particular Twitter conversations can lead to tweets with toxic language. Similarly, Habib et al. [55], utilize Perspective to identify opportunities for proactive interventions on Reddit before large escalations. Kumar et al. [89] finally determine how different types of users interact with Reddit comments labeled by the Perspective API, finding that different social groups (*e.g.*, women, racial minorities), often have different experiences when encountering the same comments.

While the Perspective API has been utilized in a host of different recent studies [57, 76, 88, 89, 124, 130] likely because of its widespread adoption by large companies like Google, Disqus, Reddit [78], several other works have sought to either improve on it utilizing newer large language models or non-machine-learning approaches. Grondahle et al. [50] show that adversarial training can make models robust to adversarial attacks like homoglyphs. Lees et al. [94] utilize a character-based transformer to build a state-of-the-art multilingual toxicity classifier that incorporates a learnable tokenizer, allowing it to be robust to domains different from its training data. Kumar et al. test recent large language models like GPT-4, Llama3, and Google Gemini, finding that they can account for ecosystems' norms and values when performing moderation [88]. In contrast to these machine-learning approaches, Jhaver et al. [77] illustrate the usefulness of the blocklists in better user experiences online. Chandrasekharan et al. [19] propose a cross-community learning strategy to build models to help moderators on Reddit detect new context content. Finally, Lai et al. [91] propose human-AI collaboration in detecting and removing content.

2.6 Present Work

Several works have studied how political polarization and online toxicity interact in particular political environments [8, 26, 144]. As argued by Ren et al. [122], these politically polarized toxic interactions online can be explained through the lens of Social Identity Theory [139], where users engaged in politically charged debates are likely to strengthen their identification with political groups, enhancing affective polarization. Most similar to our work, De Francisci Morales et al. [31] find that the interaction of individuals of different political orientations increased negative conversational outcomes. Similarly, for example, Chen et al. [22] utilize network analysis to find that misleading and highly politically divisive online videos lead to increased online incivility. Conversely, Rajadesingan et al. [119] find that political discussions in non-overtly political subreddits often lead to less toxic conversational outcomes.

In this work, however, rather than examining political polarization within a particular community or across one individual topic, we instead seek to understand, across thousands of politically engaged users across the political spectrum, the most prominent characteristics that predict increased toxicity. Subsequently, our LLM-based approach, which identifies larger topic conversations across the tweets of politically engaged Twitter/X users and multiple Twitter threads [6, 117, 151], then analyzes what contributes to polarized and toxic topics across political Twitter. Unlike previous approaches, which have largely relied on previously made hashtag lists or were limited to a set of particular topics [25] when analyzing the spread of topics, our approach is largely agnostic to these features, allowing us to analyze how various user and structural-level features contribute to toxicity across the Twitter platform. This approach enables us to study in a generalizable fashion

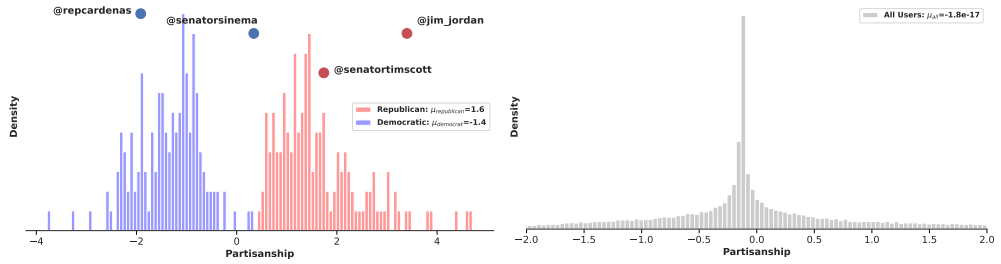


Fig. 1. Estimated Political Orientation of Political Leaders and All Users Using CA. We differentiate users' political leanings based on who they follow on Twitter.

how partisanship, polarization, and different user characteristics contribute to negative and toxic outcomes across tweets about particular subjects of varying political salience.

3 Methodology

In this section, we provide an overview of how we collected our dataset and the algorithms that we utilize to understand the interactions among Twitter users and with different topics.

3.1 Estimating Partisanship

To approximate individual Twitter users' partisanship, we rely on the Correspondence Analysis (CA) proposed by Barberá et al. [11]. Correspondence analysis (CA), similar to principal component analysis, is a technique for categorical data that extracts discriminating and representative features from a given matrix [48]. As found by Barberá et al., individual users often reveal their political preferences by whom they choose to follow on Twitter, and by analyzing these choices using CA, we can approximate their place on the political-ideological spectrum. CA works as follows: Given an $n \times m$ adjacency matrix that indicates whether user i (row) follows user j (column), CA determines a discriminating latent space among these users based on their following behaviors. By carefully choosing our initial set of "followed" users (columns of the matrix) as a set of key political figures (*e.g.*, congressional leaders), this latent space can be used to represent a dimension of "partisanship." Then, considering individuals' place on the left/right US political spectrum as a point within this latent space, we can estimate that point by projecting them onto the latent space based on who they choose to follow.[1] The result is that if a given user follows many liberal-leaning/democratic or a set of accounts that liberal-leaning accounts tend to follow, then we consider that account to be liberal [11, 102]. We note that with the CA technique, by later extending the set of key followed accounts, this approach can be used to approximate the partisanship of users who do not necessarily follow one of the initial set of key political figures (*e.g.*, congressional leaders).

We note that for our initial set of key politically predictive "followed" accounts, we utilize the Twitter accounts of the US House of Representatives and US Senate members from the 117th Congress (2021–2023). In addition to these accounts, we further add another 352 political accounts that were formerly identified by Barberá et al. (*e.g.,* @JoeBiden, @VP).[2] Using these accounts, and following the approach as specified by Barberá et al., we subsequently identified a politically ideological subspace and projected our final list of 43,151 different accounts to this subspace. See Appendix A for additional details. As seen in Figure 1, using this method, we manage to obtain a discriminating latent space that allows us to differentiate the ideology of Republican and

[1]We utilize the Tweepy API to identify the set of users that each of our non-target political accounts follows.
[2]https://github.com/pablobarbera/twitter_ideology

Democratic political leaders as well as our set of 43,151 accounts. In this setup, the more positive a user's ideology, the more right-leaning and the more negative, the more left-leaning.

3.2 Collecting Tweets

Our dataset initially consisted of 187.6M tweets from 55.4K users who followed our set of key political figures. We collected this data utilizing the Twitter API throughout 2022. We note that following the acquisition of Twitter by Elon Musk, access to the API became restricted, limiting our analysis to this time period [133]. To identify US-based users, we utilize the capability of the `Nominatim` Python tool to geocode all users' locations based on their Twitter-provided location string and OpenStreetMap.[3] Upon identifying these users, given that our work is primarily focused on the US political system, we removed any user who listed their location on their Twitter profile as outside of the United States. Altogether, we removed 12,264 users, leaving us 43,151 users. Upon identifying our user subset, we subsequently utilize `whatlanggo`[4] Go library to remove any non-English tweets from our set of users, leaving us 89,599,787 tweets. While we acknowledge several of our users' tweets might have been deleted or taken down by Twitter administrators before we scraped them, this dataset, consisting of over 89.6 million tweets, with an average of 2,076.4 (median 614.0) tweets per individual, is largely comprehensive of each user's tweeting behavior on the platform.

3.3 Determining the Toxicity of Tweets

We design and open-source[5] a contrastive learning DeBERTa-based [66] model to determine the toxicity of tweets, later benchmarking our approach on two public datasets. We further benchmark our approach against the Perspective Toxicity API [78], one of the gold standards of toxicity detection [57, 78, 87–89, 120, 142], and several other large language models. We note that for this work, we utilize encoder-based large language models to determine the toxicity of different tweets given the growing literature that has found that these types of models (*e.g.*, BERT, DeBERTa) perform better at text classification tasks (particularly within binary settings) and currently generalize better than decoder-based large language models like GPT-4o or LLaMa [59, 83, 97, 98, 114]. We note that throughout our work, we reproduce several results using the Perspective Toxicity classifier and present them in the appendix (largely obtaining similar results).

To train our new model, we rely on the Civil Comments dataset[6] that was also utilized to train and validate the Perspective API. Each comment in the dataset, depending on the percentage of 10 human raters that graded the comment as "toxic" (toxic having the definition provided in Section 2.1), is assigned a score between 0 and 1. We utilize the Civil Comments dataset given that it was specifically curated and released as part of an effort to minimize unintended bias in toxicity detection across different identity groups [35]. In addition to utilizing the training dataset of 1.8M comments, we further take two main approaches: (1) data augmentation through realistic adversarial perturbations of the original Civil Comments dataset [93], and (2) the inclusion of a contrastive learning embedding layer to help better differentiate toxic and non-toxic texts.

Specifically, for our realistic perturbations, we take advantage of the ANTHRO dataset [93]. The ANTHRO dataset consists of common online perturbations of words (*e.g.*, Republican → republiican, Reeepublican, Republicaan) extracted from online texts (*e.g.*, Twitter). For each comment with a toxicity score greater than zero in the Civil Comments training set (536,605 comments), we extract a set of random perturbations of each noun and adjective within the comment, perturbing the

[3] https://www.openstreetmap.org
[4] https://github.com/abadojack/whatlanggo
[5] The weights for our model can be downloaded at https://www.github.com/hanshanley/twits
[6] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

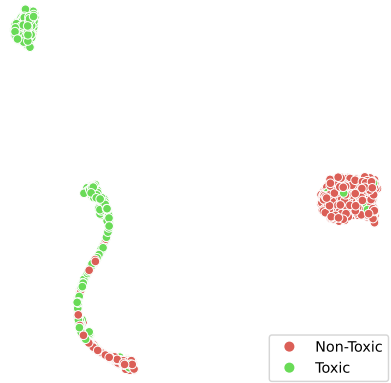


Fig. 2. **t-SNE of Civil Comments Validation Dataset** – As we train the DeBERTa-based contrastive embedding layer of our model on our augmented Civil Comments training set, our model can differentiate non-toxic (*i.e.,* toxicity $t_i < 0.5$) and toxic (*i.e.,* toxicity $t_i > 0.5$) comments. However, comments that are of ambiguous toxicity are more difficult to differentiate.

overall comment nine times with different combinations of the perturbed nouns and adjectives. This enables us to extend the set of non-zero toxicity-labeled comments to a total of 5,366,050 comments (a total of 6,634,319 toxic and non-toxic comments in the full augmented dataset). We note that in addition to allowing our model to have more training instances of toxic texts, this approach further enables our model to have training instances of real "in-the-wild" perturbations and misspellings of words that are often found on social media (*e.g.*, Twitter) and online.

For our contrastive learning objective, we train an embedding layer in a DeBERTa-based model using a supervised contrastive learning objective. This is such that while training our model, we pre-train a contrastive layer to differentiate toxic and non-toxic texts. We later freeze this layer, subsequently adding a prediction head with ReLU activation and another contextual embedding, and later training our full model to predict the toxicity of individual tweets. As shown in the 2-dimensional reduction via t-SNE [147] in Figure 2, after training this embedding layer, this layer by itself can mostly differentiate between toxic and non-toxic texts. For additional training details of our full model and our contrastive approach, see Appendix C.

Benchmarking our Toxicity Classifier. Upon training our toxicity model, we compare its performance against a vanilla fine-tuned DeBERTa model with a classification head (a two-layer MLP with ReLU activation), the widely utilized DeToxify/HateBERT [18] library[7], GPT-4o (April 11, 2025 version) [73] through the OpenAI API, the open-sourced models LLaMa 3.1 [47] and Gemma 2 [141], and the Perspective Toxicity API [78]. For the decoder-based models, we utilize the following prompt: *You are a helpful assistant who helps detect toxic online comments. Context: Toxic content is content that contains (explicit) rudeness, disrespect, or unreasonableness of a comment that is likely to make one leave the discussion. Post: "*`{row[1]}`*" Does this post contain toxic content? First, answer "Yes" or "No". Second, rate the toxicity level out of 100.* We note that while we benchmark against the GPT-4o model, given the size of our dataset (89.6M comments), for our purposes, utilizing this model for downstream classification remains impractical due to time and cost considerations.

To benchmark our toxicity model, we utilize the validation and test dataset of the Civil Comments dataset provided by Google Jigsaw [78] as well as a separate toxicity dataset provided by Kumar et al. [89]. Kumar et al.'s datasets consist of 107,620 social media comments (including from Twitter) where each comment was labeled by 5 human annotators as toxic or not (as opposed to the 10 annotators in the Civil Comments dataset). For our F_1 score calculations, as in Kumar et al. [89] and in the Civil Comments dataset, we consider a comment to be toxic if its toxicity $t_i > 0.5$. Again, we utilize this threshold for classifying a comment as toxic, given that this score (as described in

[7] https://github.com/unitaryai/detoxify

Model	CC Validation MAE	CC Validation Corr.	CC Validation Macro-F_1	CC Test MAE	CC Test Corr.	CC Test Macro-F_1	Kumar et al. MAE	Kumar et al. Corr.	Kumar et al. Macro-F_1
DeBERTa	0.0650	0.800	0.841	0.0654	0.797	0.842	0.241	0.383	0.539
DeBERTa-contrastive	**0.0601**	**0.820**	**0.851**	**0.0609**	**0.818**	**0.852**	0.251	0.415	0.540
DeToxify-original	0.0775	0.588	0.743	0.0963	0.777	0.842	0.297	0.394	0.302
DeToxify-unbiased	0.0713	0.732	0.834	0.0654	0.687	0.812	0.284	0.366	0.432
GPT-4o	0.240	0.361	0.611	0.235	0.331	0.579	0.207	**0.617**	**0.591**
LLaMa 3.1 8B (instruction-fine-tuned)	0.326	0.314	0.419	0.327	0.266	0.387	0.278	0.561	0.384
Gemma 2 7B (instruction-fine-tuned)	0.547	0.195	0.114	0.555	0.164	0.0928	**0.149**	0.617	0.499
Perspective API	0.0961	0.778	0.845	0.0963	0.777	0.842	0.332	0.417	0.410

Table 1. Mean absolute error, Pearson correlation, and F_1 score of the Perspective API and our DeBERTa models on the Civil Comments Validation and Test dataset. We bold the best scores in each respective column.

the Civil Comments task) indicates that a majority of the Civil Comments annotators would have assigned a "toxic" attribute to this comment.

As seen in Table 1, our contrastive DeBERTa model achieves the lowest mean absolute error (MAE) as well as the highest Pearson correlation and F_1 scores across the Civil Comments validation and test datasets. While our model slightly underperforms GPT-4o on the Kumar et al. dataset, it outperforms all of the other decoder-based models on this dataset and outperforms GPT-4o by a wide margin on the Civil Comments validation and test dataset. As such, for the rest of this work, when determining the toxicity of tweets, we utilize our contrastive DeBERTa model. We note that our model has a $\rho = 0.870$ Pearson correlation with the scores output by the Perspective API, illustrating its use as an offline alternative with competitive performance to Perspective. Lastly, for this work, as in other works [58, 120], when determining the overall toxicity of users or particular groupings of tweets, we utilize the average of the toxicity scores of the tweets output by our model.

3.4 Topic Analysis with MPNet and DP-Means

To later understand how particular types of users interact with different topics composed of toxic tweets, we perform topic analysis on these messages. As found by Grootendorst et al. [52, 60], by embedding small messages like tweets into a shared embedding space and then clustering these embeddings, fine-grained and highly specific topics can be extracted from datasets. To do this, we utilize the large language model MPNet[8] fine-tuned on semantic search and a parallelizable mini-batch version of the DP-Means algorithm.[9]

Fine-tuning MPNet for Topic Analysis. To compare two tweets' semantic content for later clustering, we rely on a version of the MPNet [134] large language model that was fine-tuned on semantic search. MPNet maps sentences and paragraphs to a 768-dimensional space, comparing different sentence and paragraph embeddings' semantic content based on cosine similarities (ranging from -1 [highly different] to +1 [highly similar]). We note that the version of MPNet that we utilize was initially fine-tuned on similar social media data (*e.g.*, Reddit comments, and Quora Answers), allowing us to apply this model to our set of tweets. However, to further ensure that our MPNet model is properly suited to our Twitter dataset, as in Hanley et al. [62], we further fine-tune this model using an unsupervised contrastive learning objective (*i.e.*, the SimCSE training objective) to improve the quality of our embeddings [41] on our set of tweets. As training data for this fine-tuning, we utilize 1 million tweets randomly sampled from our set of 89.6 million tweets. See Appendix B for additional details. As a reference, we provide two example tweet pairs with similarities at 0.74 and -0.03 in Figure 3. We note that for each tweet within our dataset, before embedding the message, we first remove all URLs, "@", "#", emojis, photos, and other non-textual elements from the message. In addition, for each user handle or text hashtag that utilizes camel

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[9] https://github.com/BGU-CS-VIL/pdc-dp-means

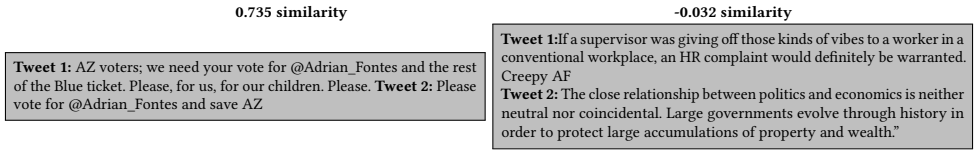


Fig. 3. Examples of Tweet pairs at different similarities (0.735 left and -0.032 right).

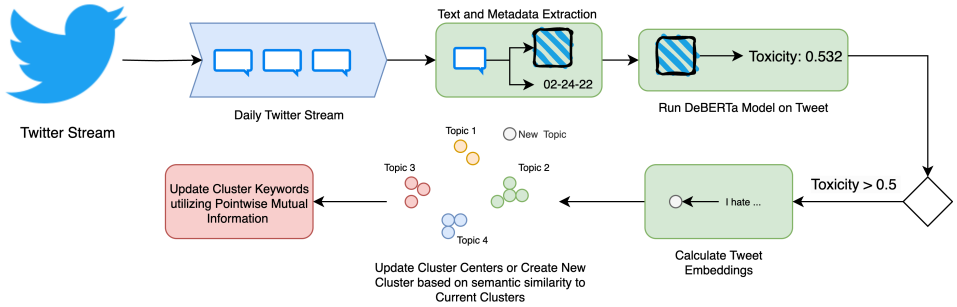


Fig. 4. Topic analysis of Toxic Tweets—We determine the toxicity, embed, and cluster toxic tweets to identify the most polarized and toxic conversations on Twitter throughout 2022. We note that for this approach, we limit our analysis to English tweets. We utilize the `whatlanggo` Go library to determine the language of tweets.

case (*i.e.*, camelCase) or snake case (*i.e.*, snake_case), we unroll those strings to their constituent elements.

DP-Means for Clustering Tweets. DP-Means [33] is a nonparametric extension of the K-means clustering algorithm. When running DP-Means, when a given datapoint is a chosen parameter λ away from the closest cluster, a new cluster is formed, and that datapoint is assigned to it; otherwise, that datapoint is assigned to its closest cluster. This characteristic of DP-Means enables us to specify how similar individual items must be to one another to be part of the same cluster. Similarly, because DP-Means is nonparametric in terms of the number of clusters formed, we do not need to know *a priori* how many topics are present within our dataset. For additional details about DP-Means, see Appendix E.

Topic Analysis Pipeline. Having outlined the constituent elements of our topic analysis algorithm, we now go over the full topic analysis pipeline (Figure 4): Throughout 2022, as we gathered the tweets of our set of 43,151 Twitter users, using our DeBERTa-contrastive model, we identify potentially toxic tweets (*i.e.*, toxicity $t_i > 0.50$). Following the identification of these potentially toxic tweets and separating out non-English tweets with `whatlanggo`, using MPNet, we subsequently map these tweets to a shared embedding space. Finally, we continuously cluster these tweets to identify topics amongst these toxic tweets using the DP-Means algorithm. To make these clusters that represent topics amongst our set of tweets, human-understandable, we employ two different approaches. First, we designate the tweets closest (*i.e.*, with the largest cosine similarity) to the center of the cluster as the "representative tweet" of the cluster [52]. Second, we determine the most distinctive keywords of each cluster using pointwise mutual information [16] (detailed in Appendix D). In this way, after clustering our set of tweets, we can later extract the semantic meaning of the various clusters outputted.

As recommended by Hanley et al., we utilize a λ of 0.60 for our clusters (precision near 0.989 for MPNet [52, 61, 62]). Finally, we extract keywords from these clusters using the pointwise mutual

information metric and determine the most representative tweets by determining the tweet with the highest cosine similarity to the cluster center. Altogether, across the 5,509,042 English-language toxic tweets from our set of 43,151 Twitter users, we identified 5,288 clusters with at least 50 toxic tweets.

3.5 Generalized Additive Models

Throughout this work, we utilize Generalized Additive Models (GAM) [65] to determine the relationships between our variables of interest (*e.g.*, user partisanship, and user toxicity). For GAMs, the relationship between the independent and dependent variables is not assumed to be linear but is rather estimated as a smooth, regularized, nonparametric function. Namely, given a dependent variable Y and a set of p independent variables X:

$$g(E(Y)) = \alpha + s_1(x_1) + \cdots + s_p(x_p), \tag{1}$$

where $g()$ is a linking function that connects the expected value of the dependent variable Y to the values of functions $s_i()$ of independent variables in X. For example, when estimating probabilities, the logit function is often utilized as with ordinary Generalized Linear Models [32]. The functions $s_i()$ represent smooth nonparametric functions of the variables in X that are fully determined by the data in X rather than by a parametric function. For GAMs, these $s_i()$ are estimated simultaneously, and the estimated value of $g(E(Y))$ is determined by adding up the values of the $s_i()$ functions. Throughout this work, we utilize the Python `pyGAM` library to fit our regressions and utilize the Generalized Cross-Validation Loss Criterion (GCV) [32] for estimating the $s_i()$ functions when fitting. The Generalized Cross-Validation Loss Criterion takes an LOOCV (Leave-One-Out Cross-Validation) approach to fitting smoothers on the data in X.

Utilizing GAMs versus other more traditional models allows us (1) to not assume linear relationships between our dependent and independent variables, and (2) to have better interpretability given that the partial contribution of a given variable x_i to determining the value of the dependent variable Y is a function only of its corresponding function $s_i()$.

3.6 Ethical Considerations

Within this work, we largely focus on identifying large-scale trends in how different Twitter users interact with one another. While we do calculate toxicity and polarization levels for individual users, we only display the names of verified public users or users with more than 500K followers, redacting the names of all other accounts. We further note that besides these public accounts, we do not publish other accounts' usernames and we do not attempt to contact nor deanonymize them. We further note that we do not analyze nor do we report on the precise behavior of individual users; instead, we only report aggregate statistics and trends in our collected data.

We note that in the training of our toxicity classifier, we utilize the Civil Comments dataset as training data, which was released in part of an effort to minimize unintended bias towards the mentioning of particular identities; for example, when the Jigsaw and Google AI teams first built toxicity classifiers, they found that the mention of the word "gay" by itself would lead to higher toxicity scores. While using this dataset helps to minimize bias, we note that our model is biased towards a consensus of toxicity that was advocated within this dataset. As found by Kumar et al. [89], individual experience of toxicity can often be highly dependent on individual users' lived experience, and any model that attempts to classify toxicity based on one standard can misrepresent these experiences. We acknowledge this flaw in our design but note that, given that within this work we are seeking to understand *general* increased levels of affective polarization among politically engaged users, personalized models largely would not work in this setting, and thus we operationalize our study using our DeBERTa-based model. Finally, we note that our Twitter

data was largely collected before Elon Musk's private acquisition of Twitter on October 27, 2022, and all of our data was collected before the later restrictions placed on the collection of tweets on June 30, 2023.[10]

4 RQ1: user level Factors in Toxicity on Twitter

Having provided background on our methodology and dataset, this section discusses several user level factors that predict and correlate with toxicity on Twitter.

4.1 Setup

Here, we examine the role of several user level factors in contributing to or affecting the rate at which individual users are toxic on Twitter. Specifically, we examine the following user characteristics in predicting the toxicity of individual users on Twitter:

(1) *The verified status of the account*
(2) *The number of years the account has been active on Twitter*
(3) *The log of the number of the account's followers*
(4) *The log of the number of accounts the user follows*
(5) *The account's partisanship as determined by our Correspondence Analysis*
(6) *The estimated average toxicity of all users the account mentioned/@ed on Twitter (i.e., accounts that the user has interacted with)*
(7) *The estimated average partisanship of the accounts the user mentioned/@ed*
(8) *The standard deviation of the partisanship of the accounts that the user mentioned (i.e., the range of political views with which the user interacts)*
(9) *The average value of the partisanship of all accounts the user mentioned/@ed*
(10) *The average difference in the partisanship of the account the given user mentioned/@ed and the user's partisanship*

We fit these ten covariates against each of our accounts' average toxicity scores. As in past studies, we fit against the verified status, the age of the account, and the information about the activity of the accounts (*e.g.*, the number of followers and the number of users followed) to understand how general account characteristics that the Twitter API returns correspond with user toxicity [21, 70]. As shown in prior work, the verification status, the number of years active, and levels of activity, depending on the context, can have differing effects on the adversarial nature and toxicity of Twitter accounts [21, 123]. Similarly, as shown in Saveski et al. [130] and Kraut et al. [84], many individual-level characteristics are predictive of users' toxicity as it predicts their level of familiarity with a given platform and their tendency to break norms (*e.g.*, post toxic content). Thus, as a baseline, and to help ground our study and determine how these account characteristics correlate with increased toxicity within the context of politically US-aligned account interactions, we include them in our model. In addition to these basic account attributes, we include information about each Twitter account's partisanship on the US left-right political spectrum as well as information about how that Twitter user interacts with other US politically aligned Twitter accounts [100]. These variables' inclusion allows us to answer our research question about whether and how affective polarization and partisanship affect the toxicity of individual accounts [85].

To understand how these factors interact with and contribute to toxicity on Twitter, we fit a Generalized Additive Model (GAM) on the average toxicity score of users (Table 2). When fitting our model, we perform variable selection using forward selection based on the Akaike Information Criterion [1], which ended up eliminating the number of followed accounts as well as user partisanship as variables from our final model. Furthermore, to ensure that our model

[10] https://help.twitter.com/en/rules-and-policies/twitter-limits

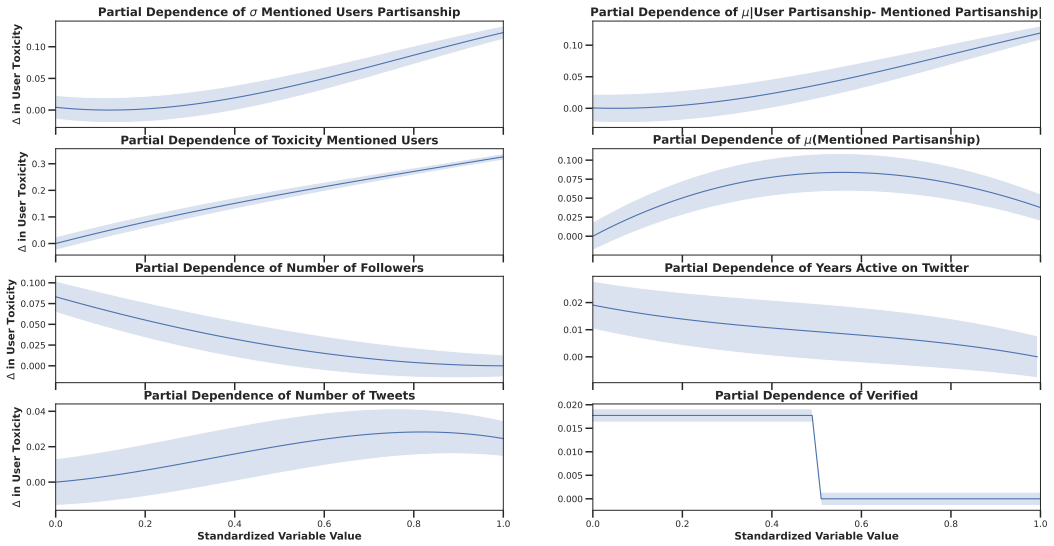


Fig. 5. Partial dependencies with 95% Normal confidence intervals between our fitted standardized dependent variables and user toxicity.

Train R^2: 0.239 , Validation R^2: 0.207			
Dependent Variable	Pearson Corr. ρ	Kendall's τ	Permut Import.
Verified Status	—	-0.242	0.053
Years Active on Twitter	-0.197	-0.147	0.027
Log # Followers	-0.206	-0.122	0.205
Log # Followed	-0.135	-0.090	—
Log # Tweets in 2022	0.147	0.200	0.045
Toxicity of Mentioned Users	0.318	0.362	0.374
Partisanship	0.054	0.061	—
σ(Mentioned partisanship)	0.317	0.332	0.150
μ(Mentioned partisanship)	0.110	0.099	0.067
μ|User partisanship- Mentioned partisanship|	0.287	0.283	0.080

Table 2. Pearson correlation ρ, Kendall's τ, and the permutation importance of dependent variables and users' toxicities. As seen in the above table, a user's interaction with a politically wide variety of users and interacting with other users with higher toxicity correlates with a given user's toxicity.

generalizes, we further reserve 10% of our data as validation, and in our results report our model's R^2 value on this validation set. Finally, after fitting this regression, we further determine the estimated importance of each variable to our final model by permuting the features and seeing the estimated impact on the R^2 score on the validation set of our data (permutation importance is a widely used statistic for determining the relative information of features to models [2]). We present the partial dependence (with 95% Normal confidence intervals) on the user toxicity of each independent variable in Figure 5 and present Pearson correlation, Kendall's τ (a more robust version of the Spearman Correlation), and each independent variable's permutation importance in Table 2. Our final model achieved an R^2 value of 0.239 on our training data and a R^2 value of 0.207 on our validation dataset, illustrating that our model does generalize to users outside of its training data.

Finally, we note that to ensure the robustness of our approach, we separately perform the same analysis utilizing the toxicity scores output by the Perspective API, obtaining similar results. We present these results in Appendix F.

4.2 Baseline Account Characteristics

We first provide an overview of how several baseline account characteristics contribute to the toxicity of each user. As seen in Table 2, we do indeed observe that each of the user characteristics that we consider (to varying degrees) *does* indeed have an observed correlational effect on how toxic users' tweets tend to be. We consider each of these effects below.

Verified Status. As seen in Table 2 and Figure 5, as also found by Hua et al. [70], whether a user is verified has a modest effect on how often they post toxic tweets, with verified users being less likely to tweet harmful or toxic messages compared to non-verified users. Overall, we find that a user's verification status has a Kendall's τ of -0.242 with their users' toxicities and has a permutation importance of 0.053 in our final model. This suggests that when users become verified and their account is associated with their offline life, users tend to be less toxic. We note that we collected users' verification status before the implementation of Twitter Blue (users could pay 8 USD to become verified) in November 2022 [40].

Years Active on Twitter. As users stay on Twitter, as seen in Figure 5, we observe that they are less likely to be toxic. As argued by Rajadesingan et al. [120] in their paper on Reddit, as social media users stay longer on particular platforms and adjust to interacting with other users, they tend to be less aggressive and toxic with other users. We see a similar result here, with older users being less toxic than younger ones. Overall, we observe that the number of years that a user is active on Twitter has a Pearson correlation of $\rho = -0.197$ with their average toxicity and a permutation importance of 0.027. This accords with past research that has found that new users, who are not used to the social mores and norms of a given online community, may more frequently violate those norms and post toxic content [84].

Number of Followers. Like verified status, and as argued by Marwick et al. [101], extremely popular users are less likely overall to be toxic than users with smaller followings. These users often create friendly public personas to interact with their followers, rarely attacking other users or posting toxic content. As seen in Figure 5, we see the same: more popular users that have more followers are less likely to post toxic tweets ($\rho = -0.206$). This variable has a permutation importance of 0.205, suggesting a high relative importance in determining the toxicity of accounts.

Number of Tweets. Many accounts in our Twitter dataset post several times a day, with the median account posting 614.0 times throughout 2022, and one account posting 413,658 times. As seen in Figure 5 with a permutation importance of 0.045 and a Pearson correlation of $\rho = 0.147$, we observe that as Twitter users post more, generally their average toxicity increases. This finding reinforces past work that suggests that accounts that post excessively and that spam Twitter are more likely to be toxic [126].

4.3 Calculated Account Characteristics: Toxicity and Political Orientation

Here, we provide an overview of how the different political and toxicity measures that we calculated contribute to individual user level toxicity.

Toxicity of Mentioned Users. We find that as users interact with or mention (@ing) other users who post toxic content, they themselves are more likely to be toxic. As seen in Figure 5, the average toxicity of accounts with which a user interacts has a nearly linear relationship with the user's own toxicity with very little variation. Indeed, we find this variable to be the most important in determining a user's toxicity, with it having a permutation importance of 0.374 and a Pearson

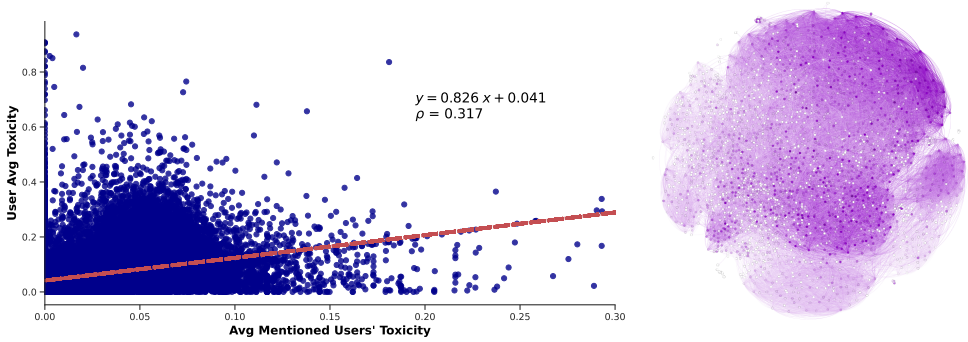


Fig. 6. The more toxic the account mentioned by a given user, on average, the more toxic the content posted by that particular user. Within the mention graph (the darker the purple, the more toxic) of user interactions, toxicity has an assortativity coefficient of 0.071, suggesting that, to some degree, users who post toxic content have a slight tendency to mention and interact with other users who post toxic content.

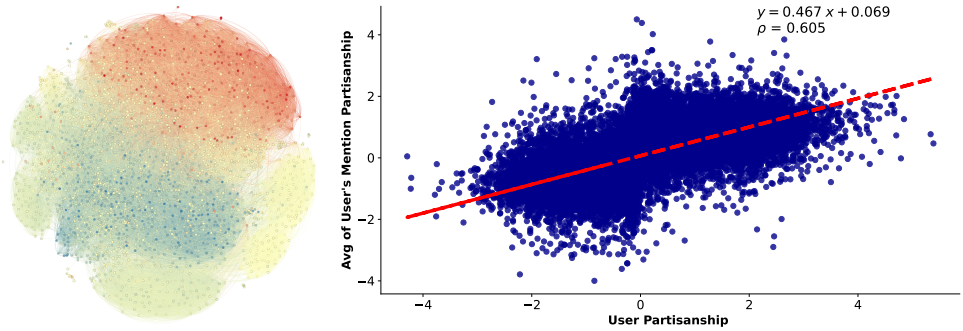


Fig. 7. Within the mention graph of user interactions (red/right-leaning and blue/left-leaning), partisanship has an assortativity coefficient of 0.266, suggesting that conservative users mention and interact more with right-leaning users while liberal users interact more with and mention other left-leaning users. Similarly, graphing the average of each user's mention's partisanship against their own partisanship, we find significant assortativity (Pearson correlation $\rho = 0.605$)

correlation $\rho = 0.318$. The most important of our covariates in terms of explainability, this result reinforces many prior findings about when and why particular users are toxic online [120, 130]. Creating a mention (@) graph among our 43,151 users and plotting users' toxicity against the toxicity of their mentioned accounts in Figure 6, we further find some degree of assortativity based on toxicity (0.071), with more toxic users more likely to interact with each other than with non-toxic users, supporting this result.

Partisanship of Mentioned Users. As the average partisanship of the accounts mentioned by a user increases (the mentioned accounts become more right-wing), we find that the average toxicity of an account increases (Figure 5) before decreasing again on the right side of the political spectrum. We thus find that when users mention users on the political extreme, this does not indicate increased toxicity; rather, we find in general that users who reference these users tend to tweet

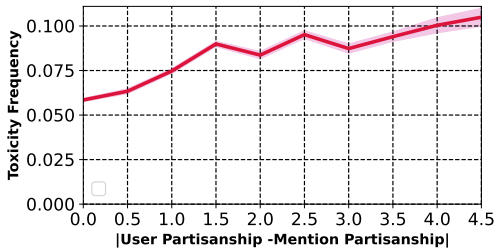


Fig. 8. As the difference in the partisanship of users and those that they mention/@ increases, the probability of users tweeting toxically increases. 95% Normal confidence Intervals.

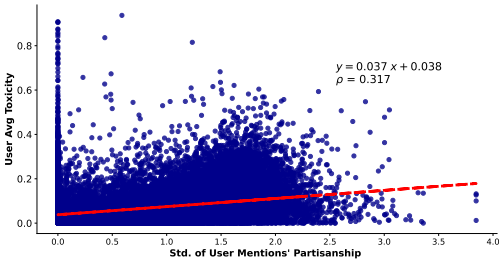


Fig. 9. As users mention a wider range of users along the political spectrum, they are more likely to tweet toxic messages.

less toxic content on Twitter. This may be due to the tendency that the users who reference these politically polarized/extreme users also tend to be near the political extremes themselves. Creating a mention/@ graph among our 43,151 users, we find a moderate degree of assortativity (0.266), thus finding that users, on the whole, tend to interact with other users of similar political views (Figure 7) and that this tendency is not necessarily correlated with increased toxicity. Graphing the average partisanship of each user's mentions against their own partisanship (Figure 7), we further observe a high assortativity (Pearson correlation of $\rho = 0.605$).

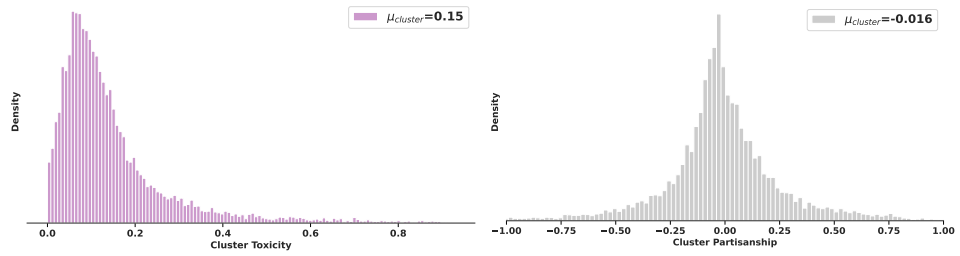
Instead, as was seen in (Figure 5), it is the difference in partisanship between a user and their mentions that linearly determines the toxicity of users. The average difference in the partisanship between a user and their mentioned accounts has a $\rho = 0.287$ Pearson correlation with the user's own toxicity and has a permutation importance of 0.080. Indeed, as seen in Figure 8, we observe across our entire dataset that as the difference between a user's partisanship and the partisanships of the corresponding users that the user mentions/@ increases, the probability that they tweet toxically also increases. This illustrates, as found elsewhere [58, 99], that as users interact with more users different in partisanship than themselves, they are more likely to be toxic. As an example, a left-wing user (-1.53) wrote the following tweet concerning the former Republican US president Donald Trump:

> *This is so indescribably fucked up. Except I love Nancy Pelosi giving him the shiv.*

Similarly, a different left-wing account (-1.504), regarding former Republican US president Donald Trump's son, wrote:

> *Fuck him. No, seriously, fuck him. If anyone's a welfare queen it's him...*

In addition to finding that as users interact with more users different than themselves, from Figure 5 and Table 2, we find that as users mention/@ a wider political diversity of users, the more toxic their own tweets. With a Pearson correlation of $\rho = 0.317$ and a permutation importance of 0.15, we see that this feature is relatively important in our fit model, with it heavily contributing to the prediction of a user's toxicity (Figure 9). This reinforces the finding of Mamakos et al. [99] who also found that when users engage with both left-leaning and right-leaning accounts on Reddit, they are more likely to engage in toxic behaviors on the platform.



(a) (b)

Fig. 10. The distribution of toxicity and partisanship within our set of clusters.

4.4 Summary

In this section, using a GAM, we explored the role that several user level characteristics play in the rate of user toxicity on Twitter. We find, most importantly, that users who interact and mention other users who regularly post toxic content are more likely to be toxic themselves. Similarly, we find that the more a given user interacts with a politically diverse set of accounts, the more likely that account is to tweet toxic content. We replicate these results with the Perspective API in Appendix F, getting similar results.

5 Factors and Changes in Polarized and Toxic Topics on Twitter

Having investigated the role that various user characteristics play in user toxicity on Twitter, we now explore how different characteristics affect different negative interactions and toxicity within conversations on Twitter. Specifically, how does the toxicity of topics on Twitter change based on the makeup of the users participating in these conversations? First, discussing and performing some qualitative analysis on the most toxic and political ideological conversations on Twitter, we then determine how the political views, the diversity of political views, and the overall toxicity of the users participating in given conversations affected particular topics discussed in 2022.

5.1 Setup

In this section, we utilize a combination of MPNet and DP-Means as specified in Section 3.4 to perform topic analysis on the English language tweets within our dataset. After running our algorithm on the 5.5M toxic tweets from our set of 43.15K Twitter users, we identified 5,288 clusters with at least 50 toxic tweets. Upon identifying these clusters, as outlined in Section 3.4, we further extract the most characteristic (often offensive) words within each cluster as well as each cluster's most representative toxic tweet. Before further detailing some of the characteristics of each of these toxic tweet clusters, we first give a brief overview of how we estimate the overall toxicity and political bent of each particular topic after identifying its corresponding cluster of toxic tweets.

Estimating the Toxicity of Topics. To estimate the toxicity of particular topics, we determine the average toxicity score of all tweets present within that given cluster. While we largely rely on our average toxicity scores, in addition to this metric, we further determine the *percentage* of toxic tweets within our *entire* English-language dataset that conform to that particular topic. Namely, after identifying each toxic cluster center, for each of these toxic cluster centers, we further identify the set of non-toxic tweets that also conform to the topic. We then calculate the percentage of toxic tweets (*i.e.*, toxicity > 0.5) per topic.

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	biden, joe, administration, president, senile	246,868	39,102 (15.84%)	0.1824	Joe Biden And everything is screwed up. You suk	0.642	0.379	1.084
2	ukraine, russia, kyiv, putin, independent	763,153	36,425 (4.77%)	0.070	So l guess you what Ukraine to stop fighting back and let the Russians kill them. Ukraine Will Resist Fuck Putin	-0.091	0.031	0.899
3	lie, pathological, truth, habitual, liar	100,825	26,894 (26.67%)	0.298	These leftist serial liars always project onto others the crimes they are perpetrating.	-0.055	0.081	1.030
4	party, democrat, republican, dnc, destroying	111,763	22,705 (20.32%)	0.232	That slate is FAR better than the gaggle of corrupt Marxists the racist lunatic democrat party pushed forward. Nobody is gonna give you a nod for badmouthing the better team.	0.215	0.171	1.162
5	ballot, election, stolen, voting, rigged	295,356	22,399 (7.58%)	0.093	You already know that the Maricopa County Election will say "Fuck Your Ballots" and ram it through the certifications.	0.199	0.121	1.131

Table 3. Top toxic topics—by the number of toxic tweets—in our dataset.

To assign non-toxic tweets to our set of toxic tweet centers, we utilize the approach laid out in prior work [60, 61] and subsequently assign each non-toxic tweet to the cluster center with the highest semantic similarity to the tweet. As recommended by Hanley et al. [62], given our fine-tuned version of MPNet, we again utilize a cluster threshold of 0.60 for assigning a given non-toxic tweet to a given cluster. We plot the distribution of estimated topic toxicity in Figure 10a. We utilize this approach, rather than clustering all 89.6 million English tweets, given the size of our dataset, and because, for this work, we are largely only concerned with topics that have some level of toxicity.

Estimating the Partisanship of Topics. To further examine the role of partisanship within interactions within particular topic clusters, we further determine the overall political orientation of each cluster. To do so, after assigning all remaining non-toxic tweets to our clusters as specified above, we subsequently determine which set of users participated in/tweeted about that topic. Calculating the average and standard deviation of the political orientations of all the Twitter users (utilizing our previous calculations of user partisanship [Section 3.1]) that tweeted about that topic, we thus estimate each topic's political-ideological composition. We plot the distribution of the partisanships of our set of clusters in Figure 10b.

5.2 The Most Toxic Topics of 2022

We start this section by providing an overview of the topics with the most toxic tweets in 2022 (Table 3). We further give an overview of the most toxic topics in Appendix G (most of these topics are merely users calling each other different epithets). As seen in Table 3, many of the most common toxic tweets concerned the most politically divisive issues of 2022 [103], namely, Joe Biden's administration (Topic 1; 247K tweets), Russia's invasion of Ukraine (Topic 2; 763K tweets),

and the abortion rights in the United States in the wake of the Dobbs v. Jackson decision which overturned US federal abortion rights [135].

Examining the average partisanship of the user who tweeted about each of the top toxic topics, we find distinct political differences. Markedly, we observe that those who tweeted in a toxic manner about the Ukraine War tended to have a slight rightward tilt (+0.031 rightward tilt). Examining these tweets, we find right-leaning users when tweeting about the war excoriated or derided the Ukrainian government or military, which was picked up as toxic by our contrastive-DeBERTa model. For example, one "toxic" tweet by a right-leaning user stated:

> *No more arms for a Ukraine refusing to negotiate! Ukraine doesn't need more arms, Ukraine needs more intelligence! And Zelensky is a dictatorial asshole!*

In contrast, considering all users who tweeted about the war, we find that they tended to lean leftward (-0.091 leftward tilt), with one left-leaning user tweeting:

> *Stand With Ukraine!*

Looking at the users who tweeted about Joe Biden's presidency (Topic 1), we again see a rightward bias (+0.642) among users who tweeted about him or his administration generally and with users who tweeted about him in a toxic manner (+0.379). For example, one user tweeted

> *Save the poor water bottle from that pedophile Joe Biden before he becomes a victim*

We thus observe that those talking about the administration (both in a toxic and non-toxic manner) were largely right-leaning (as largely expected given that the Biden administration is Democratic).

Besides these politically salient issues, we observe several topics where politically charged users simply derided each other (Topic 4) or called the other political side liars (Topic 3). We further see in Topic 5 heavy emphasis on the US presidential election being stolen in Arizona, heavily echoed by Republicans on Twitter (+0.199 rightward tilt). As documented by Prochaska et al., a misinformation story called Sharpiegate, where "Sharpies invalidated ballots in Maricopa County, Arizona" was widely spread on Twitter, and we see evidence of it in our dataset with several political users heatedly and toxically calling the Arizona election rigged [113].

5.3 Topic Dependent Changes in Partisanship and Toxicity

Having explored some of the most prominent toxic topics during our period of study, we now explore how the toxicity of different Twitter topics changed over time as users of different political orientations enter and leave. We find that regardless of whether a topic moderates (*i.e.*, political orientation moves closer to 0) or becomes more extreme (*i.e.*, political orientation becomes more left-leaning or more right-leaning), on average, this movement has little bearing on toxicity. Indeed, correlating the change in the political orientation of a given topic between January and December with the percentage change in the toxicity of that conversation, we calculate a Pearson correlation of $\rho = -0.0168$, indicating little to no relationship. Similarly, we find that the variance of political participation in particular topics over time is also only slightly correlated with the toxicity of a given topic $\rho = -0.098$. This indicates that, unlike for users, a different dynamic may be influencing the toxicity of particular topics across time.

Across our dataset, we find that regardless of whether the topic moderates or moves to the extremes, in both cases, toxicity generally increases (55.8% of the time for topics that moderated in partisanship, toxicity increases, and 71.4% of the time for topics that moved to the political extreme, toxicity increases). Furthermore, we find that between January 2022 and December 2022, in 34.8% of topics, as topics became more right-leaning, they also became more toxic; in 27.1% cases, they became less toxic as they became more right-leaning. Conversely, in 21.2% of our topics, they became more toxic as they became more left-leaning, and in 17.0% of topics, they became less toxic as they became more left-leaning. However, examining each cluster, we *do* find that on a

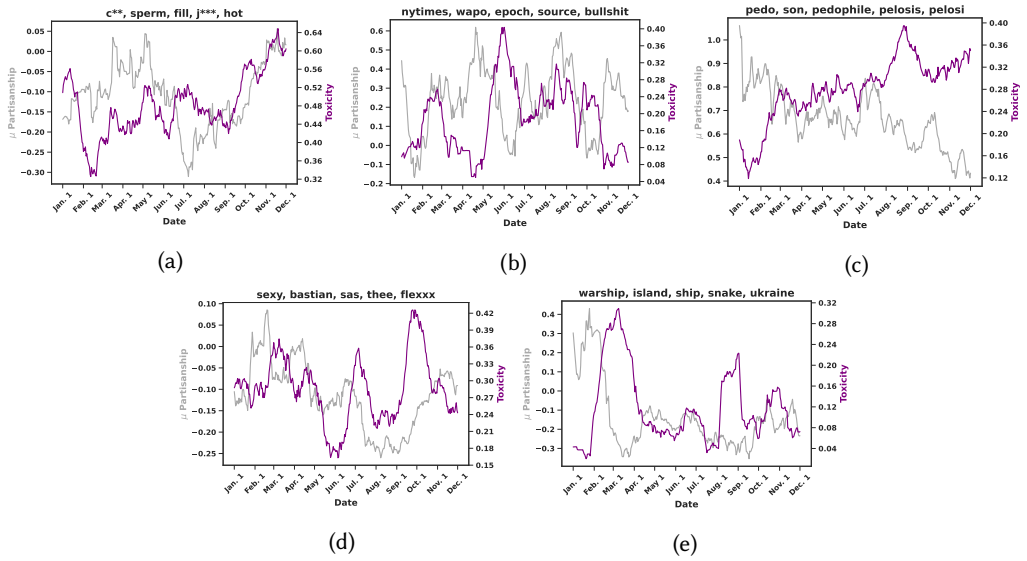


Fig. 11. Topics with the largest increase in toxicity in 2022.

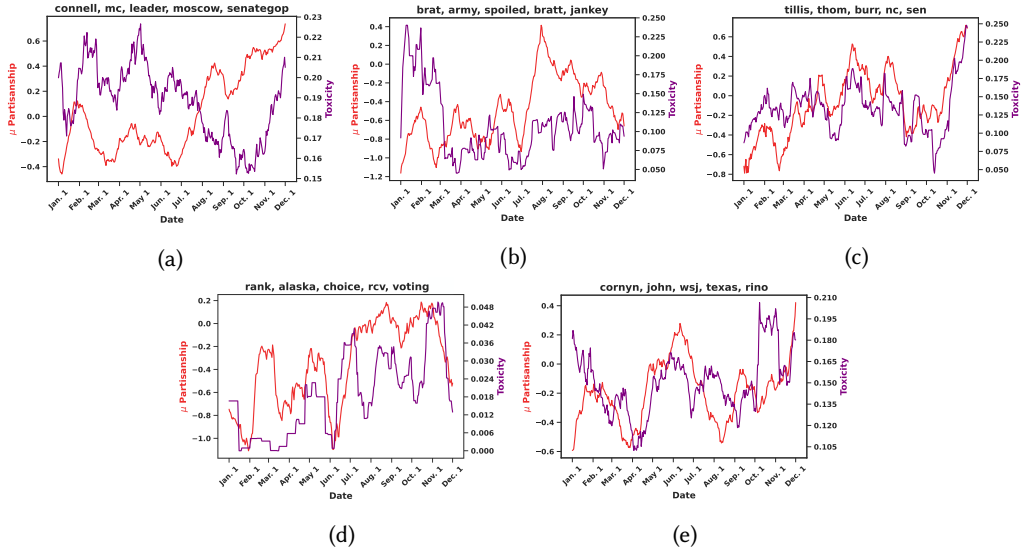


Fig. 12. Topics with the largest swing to right-leaning partisanship throughout 2022.

cluster-by-cluster basis, as the political composition of users involved in that topic changes, there are corresponding changes in toxicity.

Toxic Swings. To further qualitatively understand the nature of how toxicity and political orientation change over time, we plot the toxicity and partisanship for the topics with the largest increases in toxicity between January 2022 and December 2022. We observe that while for four topics considered, (Figures 11a, 11b, 11d, and 11e) as the topic became more right-leaning, toxicity similarly increased, for one of the topics (Figure 11c), we observe the opposite. Examining each, we

observe noticeable trends where, depending on the political nature of the topic, a corresponding swing in the political composition of the users in the left or the right direction is correlated with an increase in toxicity. For instance, we find that for the topic surrounding the destruction of the Russian warship on Snake Island by Ukrainians, the more right-wing the users became, the more toxic the surrounding conversation. For example, one user wrote:

> *Surprising Russian Navy Losses Against Ukraine Century After Tsushima Ukraine is really FUCKING Russian Navy Ship's up during the Russian Invasion into Ukraine*

In contrast, for Topic 3 (Figure 11c), we observe that as users became more left-leaning, the overall toxicity of the topic increased. We observe that this is largely due to left-leaning users adopting retorts to right-leaning users calling the Democratic former Speaker of the House, Nancy Pelosi, a pedophile. For example, we observe one user stating:

> *Let's not forget that the last republican speaker of Michigan house was a Pedophile who raped a 15 year old sister in law.*

Left-Leaning and Right-Leaning Swings. Plotting the set of topics with the largest swings in average political orientation, to both the right and left-leaning end, between January 2022 and December 2022 (Figures 12 and 13), we again observe that changes in toxicity as a result of these changes are largely dependent on the topic. For example, as the conversation surrounding Tom Tillis (the senior Republican Senator for North Carolina) became more right-leaning, the toxicity of that topic increased dramatically (Figure 12c). Despite Senator Tillis being a Republican, we observe that this is largely due to right-leaning users largely labeling Senator Tillis a RINO (Republican in name only), with one user posting:

> *You've always been a RINO NC must be ashamed of you*

We find a similar behavior for Senator John Cornyn of Texas (Figure 12e), again with a user writing:

> *John Cornyn This Bill is trash. RINOs need to go. Cornyn votes with the Democrats almost as often as his own party. Texas should be ashamed*

We similarly find that as right-leaning users joined the conversation about US Senate Republican Minority Leader Mitch McConnell being beholden to the Russian government [72], toxicity increased(Figure 12a). We note that the attacks against Senators Mitch McConnell, John Cornyn, and Tom Tillis were *all* largely for not being conservative enough. In contrast, for Republican Ohio Governor Mike DeWine (Figure 13b), we observe that as more left-leaning users joined the conversation surrounding him, the topic became more toxic, with one user writing

> *Gov Mike DeWine Thank you, Gov Mike DeWine, for making it easier for Ohioans to be killed by gun violence. Fuck you.*

Similarly, for Republican Florida Congressman Matt Gaetz (Figure 13d), we also observe that as more liberal users joined the discussions surrounding him, the topic became more toxic. We find that this was largely sparked by a tweet from Matt Gaetz stating:

> *Over-educated, under-loved millennials who sadly return from protests to a lonely microwave dinner with their cats, and no bumble matches.*

to which one user replied

> *Only stupid, insecure men worry about women being over-educated. Which one are you, matt gaetz?*

We thus observe that the context of each of these topics, in particular, is decisive for determining how different swings in political polarization will affect the overall toxicity of the topic. As for individual users (See Section 4), partisanship itself does not necessarily predict a higher degree of toxicity within conversations. Even the target/topic being a right-leaning or left-leaning entity/individual does not decisively give whether a left or right-leaning shift in users will correspond to an increase in toxicity.

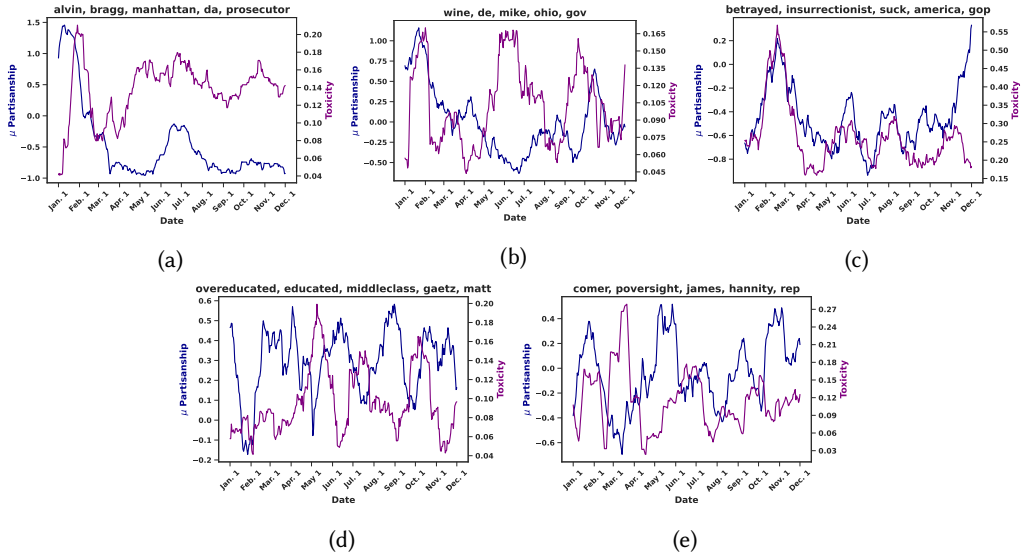


Fig. 13. Topics with the largest swing to left-leaning partisanship throughout 2022.

5.4 Topic User Composition and the Toxicity of Topics

Having qualitatively described the composition and changing dynamics of some of our set of topic clusters, we now determine how the several user level features of individual topic clusters predict the toxicity within the topic to better understand what may be influencing the toxicity of individual topics.

We note, and as seen throughout this section, topics on Twitter vary widely, with individual topics often varying widely in political composition over time. Across all topics considered in our dataset, on average between January 2022 and December 2022, the political composition of the users tweeting about each topic changed by 0.159 standard deviations (based on the latent space that we previously determined [Section 3.1]). In 61.9% of cases, topics became more right-leaning, and in 38.1% topics became more left-leaning; similarly, within this same period, 56.0% became more toxic while 44.0% became less toxic. As a result, to quantify the effect that the composition of users has on the toxicity of a given topic at a single point in time, for each topic and each month combination, we gather the user compositions and the cluster characteristic data. We thus, in this section, seek to determine the factors that predict the average toxicity score of a topic within a single month time span.

As before, to determine the role of various topic level features in the overall toxicity of that cluster, we fit a GAM on the average toxicity score each month within each of our clusters against:

(1) *The number of users who tweeted about that topic.*
(2) *The average user toxicity in the cluster.*
(3) *The percentage of users involved in that topic that is Twitter verified.*
(4) *The average of the partisanship in that cluster.*
(5) *The standard deviation of political ideologies of users within that topic cluster.*
(6) *The average age of the users in clusters.*

Again, as in Section 4, when fitting our model, we perform variable selection using forward selection based on the Akaike Information Criterion [1]. Furthermore, again, to ensure that our

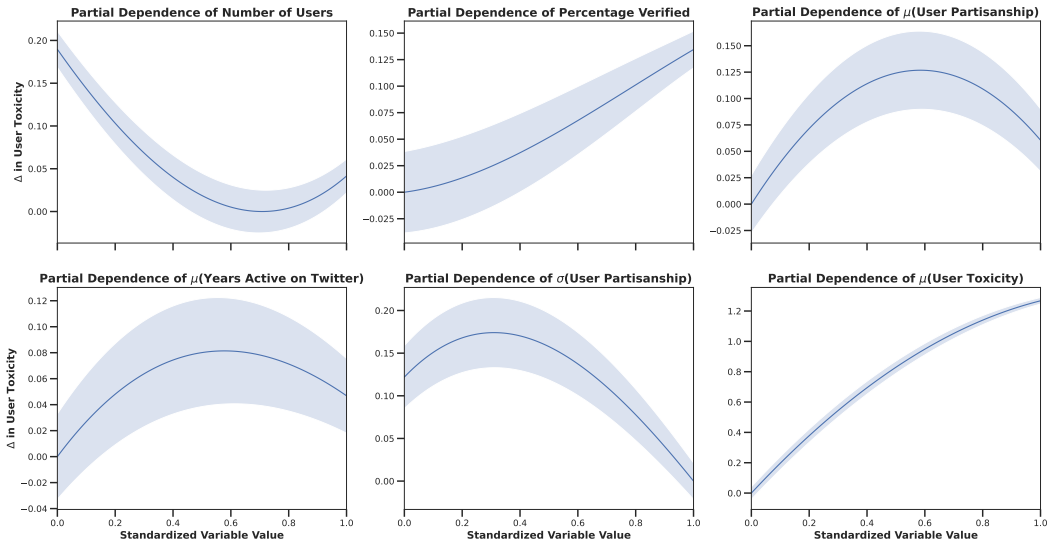


Fig. 14. Partial dependencies with 95% Normal confidence intervals between our fitted standardized dependent variables and cluster toxicity.

Train R^2: 0.397, Validation R^2: 0.389			
Dependent Variable	Pearson Corr. ρ	Kendall's τ	Permut Import.
Number of Users	-0.292	-0.139	0.445
μ(Years Active on Twitter)	-0.191	-0.186	0.018
Percentage Verified	0.234	0.250	0.007
σ(User partisanship)	0.098	0.003	0.021
μ(User partisanship)	0.028	0.005	0.007
μ(User Toxicity)	0.589	0.486	0.502

Table 4. Pearson correlation ρ, Kendall's τ, and permutation importance of dependent variables and clusters' toxicities.

model generalizes, we reserve 10% of our data as validation, and in our results report our model's R^2 value on this validation set. After fitting this regression, we further determine the estimated importance of each variable to our final model by permuting the features and seeing the estimated impact on the R^2 score of the validation set of our data. We do not consider other user account characteristics due to their multicollinearity with user toxicity (as seen in Section 4, many user characteristics are correlated with their individual toxicity). Finally, we again reproduce our results with the Perspective Toxicity API in Appendix H, obtaining similar results.

As seen in Table 4, and Figure 14, unsurprisingly, the most important factor in determining the toxicity of a given topic is the toxicity of the users contributing tweets to the cluster. This one variable has a permutation importance of 0.50 and a correlation of 0.58 with the toxicity of a given cluster. Simply put, unsurprisingly, topics whose corresponding users have higher average toxicity are more likely to have toxic content. As in Section 4, we again observe that being further along the political spectrum does not necessarily indicate increased toxicity and that a conversation being dominated by right-leaning or left-leaning users has little bearing on its toxicity.

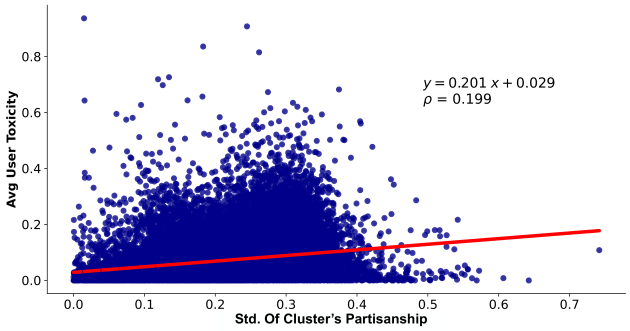


Fig. 15. As previously also found in our analysis of user characteristics (Section 4.3), we find that as users engage in a wider window of topics of particular political ideologies, the more toxic their tweeted content.

We find, as seen in Figure 14, that the number of users involved in a given topic appears to have a moderating and mitigating effect on the toxicity of that topic ($\rho = -0.292$). This also appears as one of the most important features for determining the average toxicity with a permutation score of 0.445. However, conversely, having more verified individuals participating in that topic *does* increase toxicity. We thus find (from Section 4) that while verified users are less likely to tweet toxic content, their presence and their tweeting about particular topics correlate with increased toxicity in that topic. We further find that despite the average age of accounts participating in a topic having a negative Pearson correlation with topical toxicity ($\rho = -0.191$), in our fitting model, if the average age of the accounts participating in a conversation is very young or much older, there is decreased toxicity compared to topics that engage accounts of all ages (Figure 14).

Examining political ideological contributions in Figure 14 to the toxicity of individual topics, we find that topics dominated by all left-leaning or all right-leaning users are the least toxic compared to topics in the middle of the ideological spectrum. Finally, examining the partial dependence of the diversity of viewpoints that participate in a given topic at a given point in time, we find that while initially the greater the political diversity of the topic cluster, the more toxic it becomes, as the topic invites more and more users of different beliefs that the topic cluster decreases in toxicity. While further research is needed, this result reinforces the work of Mamkos et al. [99] that finds that for particular, typically non-political topics that engage users from all over the political spectrum, these topics tend to be less toxic than others. We thus find from this analysis further confirmation, on a topic level, that increased user toxicity and the diversity of views present in a given conversation contribute to toxicity within particular topics. However, conversely, as topics invite a wider range of individuals into a discussion, toxicity actually decreases.

Lastly, looking in the reverse direction, we determine how users' toxicity changes when they are involved in many different types of politically aligned topics. As also found by Mamkos et al. [99], we find, as seen in Figure 15, as users are involved in a higher variance of topics of different political orientations, their average toxicity increases ($\rho = 0.19$). This reinforces our results from the last section as well as prior [58], which has shown that users who interact with a wider array of politically diverse users tend to be more toxic. We now consider some of the implications of these results.

6 Discussion

In this work, we determined the correlation of different aspects of partisanship and affective polarization with toxicity at a user and topic level on Twitter. We find, most notably, that users who are at the tail end of the political spectrum (very right-leaning or very left-leaning) *are not* more likely to post toxic content; rather, we observe that users that engage with a wide variety of different politically aligned accounts center have a higher likelihood of tweeting toxic messages.

Further, as users interact with or mention other users from a wider range of political ideologies, they are more likely to post toxic content. We similarly find that users who interact with other users who more regularly post toxic content are more likely to post toxic content themselves.

Examining these phenomena from a topic level, we find that most heavily partisan topics *are not* the most toxic. Rather, topics often have complex relationships with the partisanship of the users who tweet about them. While some topics become more toxic as more right-leaning/left-leaning users tweet about them, others become less toxic. However, as with individual users, we find that as users from a wider range of political ideologies tweet about a given topic, the more toxic that topic becomes. Here we discuss some of the limitations and implications of our results:

6.1 Limitations

In this work, we used a quantitative, large-scale approach to understand the interconnection between political partisanship, polarization, and toxicity at a user and topic level. We outline the limitations of our approach in this section.

Correlational Observational Study. Given our use of GAMs to estimate the effect of partisanship and political diversity and our lack of ability to perform direct experiments, our findings are largely correlational. While they do buttress and support a large literature of similar results [8, 9, 11, 29] that have found causal results in some cases for increased polarization due to interaction with users of different political beliefs, we acknowledge that *our* results are not causal. We further note that due to new restrictions placed on the collection of Tweets [133], we cannot continue to measure the toxicity of users and political topics, going forward.

Biased Dataset. Within this work, given our partisanship estimation methodology, we largely measure the interaction between US-based users and do not extend our analysis to other countries; as a result, our measurement of affective polarization in topic-wide conversations is largely limited to a US context. Similarly, even though we take pains to ensure that our dataset includes a wide range of politically active accounts (89.6M tweets across a year; all US congresspeople), we note that we did not scrape all conversations on US-based Twitter and our dataset is biased to active users that follow political or politically aligned Twitter accounts (*i.e.*, if an account did not follow any other users, we would be unable to estimate its partisanship). Finally, we note that our study, while aligning with much past work about affective polarization on social media [8, 54, 58, 108], only includes data from one year and thus likely has temporal artifacts (*e.g.*, large amounts of tweets about the Russo-Ukrainian War) that are specific to this period.

US-Based Political Study. This work largely focuses on US-based political polarization and ideologies. As a result, while applicable to dynamics for Twitter accounts on the US-political spectrum, our results do not necessarily apply to political conversations in different contexts. Similarly, we largely look at political partisanship along a unidimensional axis given the two-party system within the US. Again, as noted earlier, while this limits our analysis, given the variety of political views within the US, as found by Poole and Rosenthal, most of the variation in US political ideology is along a unidimensional axis [112] and has been utilized to much success throughout the literature [10, 63, 149]. However, given access to Twitter or similar social media websites such as Meta's Threads, our study can perhaps be replicated in different cultural contexts.

Toxicity Measurement. As found early in our work in Section 3.3, different individuals and datasets have different metrics for toxicity. While our use of Perspective API's definition of toxicity is standard throughout the literature [64, 89, 97, 120, 130], we do base our DeBERTa-based model toxicity detection on this definition; we acknowledge that it may not take into account all perspectives on what constitutes toxic online content. We point readers to our discussion of additional

limitations of this approach as well as some of the ethical considerations of measuring user toxicity in Section 3.6.

Twitter Acquisition By Elon Musk. Finally, we note that since the design and implementation of this study, Twitter has been acquired by Elon Musk, and new API restrictions have been introduced, preventing long-term analysis of changes in political toxicity within political conversations on X. Given changes in the overall algorithm on Twitter as well as due to different users leaving the platform [7, 67, 111] since our study, we note that some of our results regarding the overall political composition of users may no longer hold. However, given that our study largely measures individual users' interactions and conversations around political topics rather than the ranking of content within individual user feeds, we argue that since Twitter has not changed how users interact on the platform, we argue that many of our results are still valid. We leave it to future work to identify a means of acquiring current X data to study these dynamics.

6.2 Tribal Tendencies, Affective Polarization, Online Toxicity, and Online Echo Chambers on Twitter

As found by others, heated political conversations often elicit toxicity as people of differing views debate and discuss their differences [127]. We find that this discourse is related to increased toxicity on Twitter. This aligns with the social theory of Tajfel and Turner and the argument of Ren et al. [122] that find that people users engage in politically charged debates strengthen their identification with political groups, enhancing affective polarization. Indeed, the political diversity of those involved in a given Twitter conversation surrounding a given topic, at least in the short form of tweets, is correlated with affective polarization and toxic content. Our study thus adds nuance to previous studies of communities that have found that like-minded users gather, create distinct communities participating in a shared culture [150] that reinforce each other's views, creating toxic echo-chambers [25, 51, 136]. While users naturally often congregate and more heavily engage with users like themselves (assortativity coefficient of 0.266), showing that some echo chambers may exist on Twitter, when users exit these chambers and engage with other users of differing political views, we observe that this tends to create user conflict [53]. This result reinforces De Francisci Morales et al.'s [31] finding that interactions among users on Reddit with different political orientations have increased negative conversational outcomes, showing that it occurs in platform-wide user interactions and discussions as well. Further, indeed across all users, we find that as they increasingly interact with users of different partisanship, the frequency of toxicity increases (Figure 8). While this feature of online conversation is not the dominant factor in engendering toxic content, with other factors like a user's previous behavior [95], the age of their account, and the toxicity of other users also contributing to online toxicity, we note that this apparent "tribal tendency" appears both on a user and topic level across Twitter and across multiple Twitter threads illustrating the robustness of this finding [31, 99].

6.3 Hyperpartisan Users and Topics

In contrast to some prior work [105], we find that users and topics that are hyperpartisan (*i.e.*, very left-leaning users or very right-leaning users) are not necessarily more toxic than less ideological users. Rather, we find these users tend to mostly associate and interact with other users who share similar political views ($\rho = 0.605$) and, as a result, do not necessarily have higher toxicity levels. As also found by Grönlund et al. [51], because hyperpartisan users and topics often do not attract users of differing political views, we find that these users and topics tend to be less toxic than topics and users that interact with a wider range of the political spectrum (*i.e.*, topics and users nearer to the political center). This result indicates that political echo chambers, where only left-leaning or

right-leaning individuals interact among themselves, may be less conflict-oriented on Twitter. As a result, we argue that if social media companies like Twitter wish to expose their users to a wider range of political views without increasing conflict on their platforms, these users may be more amenable to these opposing views if they come from others nearer to themselves on the political spectrum.

6.4 Intra-Topic Partisanship over Time

In Section 5.3, we observed that the political orientation of users who discuss any particular topic often changes over time. These changes, often coinciding with changes in toxicity, also illustrate that the views expressed on Twitter about particular topics often change as different users enter or leave conversations. We argue that future analysis of topics and their spread on Twitter *must* take into account user level characteristics such as partisanship, given that these values often reveal the nature of how users are addressing individual topics. For example, as seen in Section 5.3, understanding that conversations surrounding "Moscow Mitch" had been taken up by increasingly right-leaning users reveals the penetration of this insult into more conservative circles.

6.5 Toxic Birds of a Feather

In addition to finding that the range of political views encountered by a particular user is predictive of toxicity, we further find that topics and users who interact with other toxic users are more likely to be toxic themselves. This again buttresses prior work from Kim et al., Kwon et al., and Shen et al. who all find that exposure to these negative conversations actually increases observers' tendency to also engage in incivility [82, 90, 132]. While not a new finding [88], this illustrates that reducing toxic content online may have other downstream benefits; by removing more instances of toxic content, other users may be less likely to engage in toxicity themselves, further reducing the amount of toxic content. Given the existence of particular toxicity norms within communities Reddit [120], where toxicity is more rarely seen among users and toxic comments are looked down upon, we argue that removing toxic content may have a compounding effect, greatly improving the overall health of online discourse.

6.6 Implications and Recommendations for the Twitter/X Platform

Our work simultaneously finds that topics that engage with a wider set of politically aligned users and users that engage in a wider array of different political discussions are more likely to tweet toxic messages. Namely, exposure on the Twitter/X platform to differing views may essentially be counterproductive to producing civil online discussions [8]. Furthermore, this suggests that recent attempts to widen the range of political discussion on Twitter may have the additional effect of increasing online toxicity [68]. As such, we argue that as Twitter continues to widen the political conversation on its platform, to also maintain low levels of toxicity, additional moderation steps or additional practices should be taken to slowly introduce users to other accounts with different political beliefs from themselves, should be taken as well [104]. Practically, this could involve down-weighting political content that has the polar opposite views of the users and slightly up-ranking political content that is only somewhat dissimilar to the user. This accords with the recommendations and findings of Mamakos et al. [99] who found that as Reddit users engage with users different from them and in a wider variety of political contexts, they tend to be more toxic. Given that Twitter users are not siphoned out into individual communities that they specifically join and thus more easily engage with polarizing content and users with whom they disagree across their topics of interest, we argue that building a means by which to engage in better conversations across political differences can reduce toxicity and friction on the platform. For example, as also argued by [105], including a wide and generalized view of particular topics could potentially reduce

polarization. Indeed, as found in Section 5.4, while initially sparking more toxicity, as topics include a wider and wider berth of political perspectives and as more users join a topic, the toxicity of that particular topic decreases.

As shown elsewhere [64, 78, 87, 97], ML-based approaches can be utilized to help track toxic behavior online, and we finally note that our open-source DeBERTa-based model, combined with our approach of mapping user partisanship, can further assist in helping identify particularly politically charged conversations online that have devolved. As shown throughout Section 5, by examining changes in user political composition and toxicity over time, we can identify large changes in toxicity surrounding particular topics as well as which set of users is driving this change. We propose that by creating a dashboard to identify which topics are driving toxicity, platforms can take action on particular platform-wide conversations that degrade the quality of interactions between users. By understanding these dynamics in real time and knowing which topics are driving toxicity, we argue that platforms can take steps to help mitigate the most divisive and politically toxic conversations on their platforms. For example, by down-ranking particular divisive and toxic conversations that are quickly being hijacked by particular users, platforms can help ameliorate the spread of toxic posts [12]. Further, given that we utilize a relatively smaller DeBERTa-based and open-sourced model that can be run locally on a commodity GPU (*e.g.*, NVIDIA RTX A6000), we note that it can be scaled to large datasets much more easily than other API or decoder-LLM-based approaches.

6.7 Future Work

This work centered around understanding factors that contribute to the toxicity levels of individual users and within particular topics on Twitter/X. However, we note that several of the techniques employed within this work can be extended and utilized beyond our study.

Identifying the Role of Partisanship and Polarization on Different Platforms In this work, while we focus on Twitter, we note that our approach can largely be utilized on different social media platforms (*e.g.*, Facebook, Reddit, Bluesky, Threads, *etc...*) to identify the role of partisanship and political polarization. Unlike on Twitter, where a feed is curated for the user, Reddit user interactions, for instance, are largely determined by the communities into which the user self-selects. Previous work has shown that entire communities can engage in cross-partisan toxic behavior [37]. Similarly, Bail et al. [8] find that simply following users and repeatedly seeing disagreeable content can increase polarization. As such, we plan to explore the robustness of our findings about "tribal tendencies" in different contexts and what best practices can be utilized to ameliorate these tendencies.

7 Conclusion

In this work, we analyzed which factors predict toxicity at the user and the topic level on Twitter. We propose, implement, and release a new open-source toxicity classifier, achieving better accuracy than the Perspective API and many state-of-the-art decoder-based large language models on the Civil Comments dataset. Then, analyzing 89.6M tweets posted by 43.15K users from across the political spectrum, we find that a user or topic being heavily partisan does not necessarily imply increased toxicity; rather, as users engage with and have conversations involving a wider range of political orientations and with other toxic users that their own online toxicity increases. We recommend that platforms, given these findings, take pains to ensure that users, while not put in politically homogeneous echo chambers, be *slowly* introduced to other accounts with different political orientations from themselves [104].

Acknowledgments

This work was supported in part by the NSF Graduate Fellowship DGE-1656518, a Meta Ph.D. Fellowship, and a Sloan Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or other funding agencies.

References

[1] Hirotugu Akaike. 2011. Akaike's information criterion. *International encyclopedia of statistical science* (2011), 25–25.
[2] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (2010), 1340–1347.
[3] Jisun An, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2011. Media landscape in Twitter: A world of new conventions and political diversity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 18–25.
[4] Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 68–79.
[5] Jisun An, Daniele Quercia, and Jon Crowcroft. 2014. Partisan sharing: Facebook evidence and societal consequences. In *Proceedings of the second ACM conference on Online social networks*. 13–24.
[6] Okan Arslan, Wanli Xing, Fethi A Inan, and Hanxiang Du. 2022. Understanding topic duration in Twitter learning communities using data mining. *Journal of Computer Assisted Learning* 38, 2 (2022), 513–525.
[7] Arvindh Arun, Saurav Chhatani, Jisun An, and Ponnurangam Kumaraguru. 2024. X-posing Free Speech: Examining the Impact of Moderation Relaxation on Online Social Networks. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*. 201–211.
[8] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.
[9] Pablo Barberá. 2014. How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. *Job Market Paper, New York University* 46 (2014), 1–46.
[10] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.
[11] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.
[12] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, et al. 2023. Embedding societal values into social media algorithms. *Journal of Online Trust and Safety* 2, 1 (2023).
[13] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. 2016. Users polarization on Facebook and Youtube. *PloS one* 11, 8 (2016), e0159641.
[14] Porismita Borah. 2013. Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication* 30, 3 (2013), 456–473.
[15] Mark Boukes. 2019. Social network sites and acquiring current affairs knowledge: The impact of Twitter and Facebook usage on learning about the news. *Journal of Information Technology & Politics* 16, 1 (2019), 36–51.
[16] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009).
[17] Pamela Jo Brubaker, Daniel Montez, and Scott Haden Church. 2021. The power of schadenfreude: Predicting behaviors and perceptions of trolling among Reddit users. *Social Media+ Society* 7, 2 (2021), 20563051211021382.
[18] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472* (2020).
[19] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.
[20] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.
[21] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring# GamerGate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international*

conference on world wide web companion. 1285–1290.
[22] Yingying Chen and Luping Wang. 2022. Misleading political advertising fuels incivility online: A social network analysis of 2020 US presidential election campaign video comments on YouTube. *Computers in Human Behavior* 131 (2022), 107202.
[23] Shira Chess and Adrienne Shaw. 2015. A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media* 59, 1 (2015), 208–220.
[24] Jaeho Cho, Saifuddin Ahmed, Martin Hilbert, Billy Liu, and Jonathan Luu. 2020. Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media* 64, 2 (2020), 150–172.
[25] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.
[26] Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. Dynamics of online hate and misinformation. *Scientific reports* 11, 1 (2021), 1–12.
[27] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 192–199.
[28] Dana Cuomo and Natalie Dolci. 2019. Gender-based violence and technology-enabled coercive control in Seattle: Challenges & Opportunities. *TECC Whitepaper Series* (2019).
[29] Chrysi Dagoula. 2019. Mapping political discussions on Twitter: Where the elites remain elites. *Media and Communication* 7, 1 (2019), 225–234.
[30] Sheila Dang, Kenneth Li, and Matthew Lewis. 2022. Exclusive: Twitter is losing its most active users, internal documents show | Reuters. https://www.reuters.com/technology/exclusive-where-did-tweeters-go-twitter-is-losing-its-most-active-users-internal-2022-10-25/
[31] Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. 2021. No echo in the chambers of political interactions on Reddit. *Scientific reports* 11, 1 (2021), 1–12.
[32] Alfred Demaris. 1992. *Logit modeling: Practical applications*. Number 86. Sage.
[33] Or Dinari and Oren Freifeld. 2022. Revisiting DP-Means: Fast Scalable Algorithms via Parallelism and Delayed Cluster Creation. In *The 38th Conference on Uncertainty in Artificial Intelligence*.
[34] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. Affective polarization, local contexts and public opinion in America. *Nature human behaviour* 5, 1 (2021), 28–38.
[35] Corentin Duchene, Henri Jamet, Pierre Guillaume, and Reda Dehak. 2023. A benchmark for toxic comment classification on civil comments dataset. *arXiv preprint arXiv:2301.11125* (2023).
[36] Maeve Duggan. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.
[37] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2023. Non-polar opposites: analyzing the relationship between echo chambers and hostile intergroup interactions on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 197–208.
[38] Claudia Flores-Saviaga, Brian Keegan, and Saiph Savage. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[39] Jasmine C Foriest, Shravika Mittal, Kirsten Bray, Anh-Ton Tran, and Munmun De Choudhury. 2024. A Cross Community Comparison of Muting in Conversations of Gendered Violence on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–29.
[40] Brian Fung. 2023. How Elon Musk transformed Twitter's blue check from status symbol into a badge of shame | CNN Business. https://www.cnn.com/2023/04/24/tech/musk-twitter-blue-check-mark/index.html
[41] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552
[42] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.
[43] Anthony J Gaughan. 2016. Illiberal democracy: The toxic mix of fake news, hyperpolarization, and partisan election administration. *Duke J. Const. L. & Pub. Pol'y* 12 (2016), 57.
[44] Bryan T Gervais. 2015. Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics* 12, 2 (2015), 167–185.

[45] Vahid Ghafouri, Faisal Alatawi, Mansooreh Karami, Jose Such, and Guillermo Suarez-Tangil. 2024. Transformer-Based Quantification of the Echo Chamber Effect in Online Communities. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–27.

[46] Ine Goovaerts and Sofie Marien. 2020. Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication* 37, 6 (2020), 768–788.

[47] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[48] Michael J Greenacre. 2010. Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 5 (2010), 613–619.

[49] Kirsikka Grön and Matti Nelimarkka. 2020. Party Politics, Values and the Design of Social Media Services: Implications of political elites' values and ideologies to mitigating of political polarisation through design. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–29.

[50] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

[51] Kimmo Grönlund, Kaisa Herne, and Maija Setälä. 2015. Does enclave deliberation polarize opinions? *Political Behavior* 37 (2015), 995–1020.

[52] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[53] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. Avoiding the echo chamber about echo chambers. *Knight Foundation* 2, 1 (2018), 1–25.

[54] Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* 381, 6656 (2023), 398–404.

[55] Hussam Habib, Maaz Bin Musa, Muhammad Fareed Zaffar, and Rishab Nithyanand. 2022. Are Proactive Interventions for Reddit Communities Feasible?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 264–274.

[56] Yosh Halberstam and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* 143 (2016), 73–88.

[57] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric. 2024. PressProtect: Helping Journalists Navigate Social Media in the Face of Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–34.

[58] Hans WA Hanley and Zakir Durumeric. 2023. Sub-Standards and Mal-Practices: Misinformation's Role in Insular, Polarized, and Toxic Interactions. *arXiv preprint arXiv:2301.11486* (2023).

[59] Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 542–556.

[60] Hans WA Hanley and Zakir Durumeric. 2024. Partial mobilization: Tracking multilingual information flows amongst russian media outlets and telegram. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 528–541.

[61] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2023. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (2023).

[62] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2024. Specious sites: Tracking the spread and sway of spurious news stories at scale. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1609–1627.

[63] Hans WA Hanley, Emily Okabe, and Zakir Durumeric. 2025. Tracking the Takes and Trajectories of English-Language News Narratives across Trustworthy and Worrisome Websites. In *34th USENIX Security Symposium (USENIX Security 25)*.

[64] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3309–3326.

[65] Trevor J Hastie. 2017. Generalized additive models. In *Statistical models in S*. Routledge, 249–307.

[66] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

[67] Daniel Hickey, Daniel MT Fessler, Kristina Lerman, and Keith Burghardt. 2025. X under Musk's leadership: Substantial hate and no reduction in inauthentic activity. *PloS one* 20, 2 (2025), e0313293.

[68] Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E Smaldino, Goran Muric, and Keith Burghardt. 2023. Auditing Elon Musk's impact on hate speech and bots. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17. 1133–1137.
[69] Sounman Hong and Sun Hyoung Kim. 2016. Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly* 33, 4 (2016), 777–782.
[70] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[71] Robert Huckfeldt, Paul Allen Beck, Russell J Dalton, and Jeffrey Levine. 1995. Political environments, cohesive social groups, and the communication of public opinion. *American Journal of Political Science* (1995), 1025–1054.
[72] Carl Hulse. 2018. 'Moscow Mitch' Tag Enrages McConnell and Squeezes G.O.P. on Election Security - The New York Times. https://www.nytimes.com/2019/07/30/us/politics/moscow-mitch-mcconnell.html
[73] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
[74] Roland Imhoff, Felix Zimmer, Olivier Klein, João HC António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, et al. 2022. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour* 6, 3 (2022), 392–403.
[75] Irina Ivanova. 2023. Twitter is now X. Here's what that means. https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means/
[76] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. 2018. Adversarial text generation for google's perspective api. In *2018 international conference on computational science and computational intelligence (CSCI)*. IEEE, 1136–1141.
[77] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.
[78] Google Jigsaw. 2022. Google Jigsaw. Perspective API. https://www.perspectiveapi.com/home.
[79] Andreas Jungherr. 2014. Twitter in politics: a comprehensive literature review. *Available at SSRN 2402443* (2014).
[80] Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph. D. Dissertation.
[81] Amir Karami, Morgan Lundy, Frank Webb, and Yogesh K Dwivedi. 2020. Twitter and research: A systematic literature review through text mining. *IEEE Access* 8 (2020), 67698–67717.
[82] Yonghwan Kim and Youngju Kim. 2019. Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion. *Computers in Human Behavior* 99 (2019), 219–227.
[83] Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large Language Models For Text Classification: Case Study And Comprehensive Review. *arXiv preprint arXiv:2501.08457* (2025).
[84] Robert Kraut, Moira Burke, John Riedl, and Paul Resnick. 2010. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities* 1 (2010), 42.
[85] Emily Kubin and Christian Von Sikorski. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* 45, 3 (2021), 188–206.
[86] Brian Kulis and Michael I Jordan. 2011. Revisiting k-means: New algorithms via Bayesian nonparametrics. *arXiv:1111.0352* (2011).
[87] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 865–878.
[88] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*. 2797–2807.
[89] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 299–318.
[90] K Hazel Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* (2017).
[91] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[92] Huyen T Le, GR Boynton, Yelena Mejova, Zubair Shafiq, and Padmini Srinivasan. 2017. Revisiting the american voter on twitter. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 4507–4519.
[93] Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu, and Dongwon Lee. 2022. Perturbations in the Wild: Leveraging Human-Written Text Perturbations for Realistic Adversarial Attack and Defense. *60th Annual Meeting of the Association for Computational Linguistics (ACL)* (2022).

[94] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3197–3207.
[95] Sharon Levy, Robert E Kraut, Jane A Yu, Kristen M Altenburger, and Yi-Chia Wang. 2022. Understanding Conflicts in Online Conversations. In *Proceedings of the ACM Web Conference 2022*. 2592–2602.
[96] Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. Association for Computational Linguistics, 81–91.
[97] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389* (2023).
[98] Mahdi Maktabdar Oghaz, Lakshmi Babu Saheer, Kshipra Dhame, and Gayathri Singaram. 2023. Detection and classification of ChatGPT generated contents using deep transformer models. *Frontiers in Artificial Intelligence* 8 (2023), 1458707.
[99] Michalis Mamakos and Eli J Finkel. 2023. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS nexus* 2, 10 (2023), pgad325.
[100] Fabrizio Marozzo and Alessandro Bessi. 2018. Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining* 8 (2018), 1–13.
[101] Alice Marwick and Danah Boyd. 2011. To see and be seen: Celebrity practice on Twitter. *Convergence* 17, 2 (2011), 139–158.
[102] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001), 415–444.
[103] Domenico Montanaro. 2022. Abortion, inflation, Ukraine: 2022's top U.S. political stories : NPR. https://www.npr.org/2022/12/31/1146261338/2022-political-stories-midterms-abortion-inflation-immigration-ukraine
[104] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
[105] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization. In *Proceedings of the 2018 designing interactive systems conference*. 957–970.
[106] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. 2017. Measuring offensive speech in online political discourse. In *7th USENIX workshop on free and open communications on the internet (FOCI 17)*.
[107] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.
[108] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620, 7972 (2023), 137–144.
[109] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.
[110] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.
[111] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, et al. 2023. Just another day on Twitter: a complete 24 hours of Twitter data. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17. 1073–1081.
[112] Keith T Poole and Howard Rosenthal. 2007. On party polarization in Congress. *Daedalus* 136, 3 (2007), 104–107.
[113] Stephen Prochaska, Kayla Duskin, Zarine Kharazian, Carly Minow, Stephanie Blucker, Sylvie Venuto, Jevin D West, and Kate Starbird. 2023. Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 US presidential election. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–39.
[114] Yunjian Qiu and Yan Jin. 2024. ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent systems with applications* 21 (2024), 200308.
[115] Walter Quattrociocchi, Rosaria Conte, and Elena Lodi. 2011. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems* 14, 04 (2011), 567–586.
[116] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).
[117] Daniele Quercia, Licia Capra, and Jon Crowcroft. 2012. The social world of twitter: Topics, geography, and emotions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. 298–305.
[118] Stephen A Rains, Kate Kenski, Kevin Coe, and Jake Harwood. 2017. Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*

22, 4 (2017), 163–178.
[119] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. 2021. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, Vol. 15.
[120] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 557–568.
[121] Kunal Relia, Zhengyi Li, Stephanie H Cook, and Rumi Chunara. 2019. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 417–427.
[122] Yuqing Ren, Robert Kraut, and Sara Kiesler. 2007. Applying common identity and bond theory to design of online communities. *Organization studies* 28, 3 (2007), 377–408.
[123] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[124] Bernhard Rieder and Yarden Skop. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 8, 2 (2021), 20539517211046181.
[125] Jon C Rogowski and Joseph L Sutherland. 2016. How ideology fuels affective polarization. *Political behavior* 38 (2016), 485–508.
[126] Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User Engagement and the Toxicity of Tweets. *arXiv preprint arXiv:2211.03856* (2022).
[127] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PloS one* 15, 2 (2020), e0228723.
[128] Martin Saveski, Doug Beeferman, David McClure, and Deb Roy. 2022. Engaging Politically Diverse Audiences on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 873–884.
[129] Martin Saveski, Nabeel Gillani, Ann Yuan, Prashanth Vijayaraghavan, and Deb Roy. 2022. Perspective-taking to reduce affective polarization on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 885–895.
[130] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*. 1086–1097.
[131] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences* 114, 12 (2017), 3035–3039.
[132] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (2020), 106343.
[133] Manish Singh. 2023. Twitter limits the number of tweets users can read amid extended outage | TechCrunch. https://techcrunch.com/2023/07/01/twitter-imposes-limits-on-the-number-of-tweets-users-can-read-amid-extended-outage/
[134] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Adv. in Neural Information Processing Systems* (2020).
[135] Washington Post Staff. 2022. Read the full opinion in Dobbs v. Jackson Women's Health O - Washington Post. https://www.washingtonpost.com/politics/interactive/2022/roe-wade-decision-pdf/
[136] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. Ecosystem or echo-system? Exploring content sharing across alternative media domains. In *Proceedings of the International AAAI Conference on Web and Social Media*.
[137] Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. 2018. The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics* 23, 1 (2018), 95–115.
[138] Cass R Sunstein. 2018. Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.* 27 (2018), 83.
[139] Henri Tajfel and John C Turner. 1978. Intergroup behavior. *Introducing social psychology* 401, 466 (1978), 149–178.
[140] Edson C Tandoc Jr and Erika Johnson. 2016. Most students get breaking news first from Twitter. *Newspaper research journal* 37, 2 (2016), 153–166.
[141] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).
[142] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.
[143] Christopher Torres-Lugo, Kai-Cheng Yang, and Filippo Menczer. 2022. The Manufacture of Partisan Echo Chambers by Follow Train Abuse on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*,

Vol. 16. 1017–1028.
[144] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).
[145] Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *Journal of democracy* 28, 4 (2017), 46–59.
[146] Peter D Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on machine learning*.
[147] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[148] Stephanie Wang, Shengchun Huang, Alvin Zhou, and Danaë Metaxa. 2024. Lower Quantity, Higher Quality: Auditing News Content and User Perceptions on Twitter/X Algorithmic versus Chronological Timelines. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–25.
[149] Galen Weld, Maria Glenski, and Tim Althoff. 2021. Political bias and factualness in news sharing across more than 100,000 online communities. In *Proceedings of the international AAAI conference on web and social media*, Vol. 15. 796–807.
[150] Etienne Wenger, Richard McDermott, and William M Snyder. 2002. Seven principles for cultivating communities of practice. *Cultivating Communities of Practice: a guide to managing knowledge* 4 (2002), 1–19.
[151] Maranke Wieringa, Daniela van Geenen, Mirko Tobias Schäfer, and Ludo Gorzeman. 2018. Political topic-communities and their framing practices in the Dutch Twittersphere. *Internet Policy Review* 7, 2 (2018), 1–16.
[152] Magdalena Wojcieszak, Andreu Casas, Xudong Yu, Jonathan Nagler, and Joshua A Tucker. 2022. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science advances* 8, 39 (2022), eabn9418.
[153] Magdalena E Wojcieszak and Diana C Mutz. 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication* 59, 1 (2009), 40–56.
[154] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.
[155] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–23.
[156] Yulin Yu, Julie Jiang, and Paramveer S Dhillon. 2024. Characterizing the Structure of Online Conversations Across Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–23.

A Correspondence Analysis for Approximating Political Ideology

After identifying our set of 882 politically discriminating accounts and identifying 6,107 random accounts that followed this set of accounts, we performed the following for CA.

(1) **Identify the Ideological Subspace:** Using 6,107 accounts that followed 10 or more of our 882 discriminating political users, we derive an initial CA model and obtain a discriminating latent space on which to plot user political ideology.
(2) **Expand the number of discriminating political ideological accounts:** Utilizing our initial CA model we determine the set of Twitter accounts not included within our initial target accounts that were most often followed by the most conservative and liberal accounts (within the top 20% on either side of the political spectrum) in the first stage of our analysis. As in Barberá et al. [11], we compute the popularity among users of a given ideological orientation such that $pop_{jc} = n_{jc} - n_{jl}$ for conservatives, where n_{jc} is the number of conservative users included in the first stage that follow account j, and n_{jl} is the equivalent measure for liberals. We further filter these accounts to ensure that at least 3 different users follow these additional discriminating accounts. After determining these users, we add the resulting 788 accounts as additional "following" accounts to our original $n \times m$ matrix. These additional accounts include those of Barack Obama (@BarackObama), MSNBC (@MSNBC), Florida governor Ron DeSantis (@GovRonDeSantis), and the House GOP (@HouseGOP).

(3) **Expanding the number of follower accounts:** For the rest of our users, we project them into the discriminating latent space utilizing our CA model. This allows us to utilize the information from our original discriminating political accounts as well as from the additional discriminating political accounts from the second stage. We can further estimate the political ideology of any account that follows at least one of 1670 highly politically discriminating accounts. After projecting all of our users, we standardize the estimates into z-scores (*i.e.,* a value of 0 represents the average partisanship and a value of 1 represents one standard deviation above the mean, 2, two standard deviations above the mean, *etc...*).

B Unsupervised Contrastive Learning

We utilize the SimCSE training objective to further refine our MPNet model and ensure that it is properly suited for our dataset. This is such that we embed each tweet i $x_i = (tweet_i) \in D_{tweets}$ (where $tweet_i$ is the text) twice (with dropout both times) using MPNet by inputting $[CLS]text_i[SEP]$ and outputting out the contextual hidden vectors $\mathbf{h}_i$ and $\tilde{\mathbf{h}}_i$ for $text_i$ as its representations. Then, given a batch of contextual hidden vectors $\{\mathbf{h}_i\}_{i=0}^{N_b}$ and $\{\tilde{\mathbf{h}}_j\}_{j=0}^{N_b}$ (different dropout), where N_b is the size of the batch, for each batch in our training dataset of 1 million tweets, we perform a contrastive learning step on that batch. This is such that for each batch $\mathcal{B}$, for an *anchor* hidden embedding $\mathbf{h_i}$ within the batch, the set of hidden contextual vectors $\mathbf{h_i}\, \tilde{\mathbf{h}}_\mathbf{j} \in \mathcal{B}$, the hidden contextual vectors where $i = j$ are positive pairs. Other pairs where $i \neq j$ are considered negative pairs. Within each batch $\mathcal{B}$, the contrastive loss is computed across all positive pairs in the batch such that:

$$L_{contrastive} = -\frac{1}{N_b} \sum_{\mathbf{h}_i \in \mathcal{B}} l^c(\mathbf{h}_i)$$

$$l^c(\mathbf{h}_i) = log \frac{\sum_{j \in \mathcal{B}} \mathbb{1}_{[i=j]} \exp\left(\frac{\mathbf{h}_i^\top \tilde{\mathbf{h}}_j}{\tau ||\mathbf{h}_i|| ||\tilde{\mathbf{h}}_j||}\right)}{\sum_{j \in \mathcal{B}} \exp\left(\frac{\mathbf{h}_i^\top \tilde{\mathbf{h}}_j}{\tau ||\mathbf{h}_i|| ||\tilde{\mathbf{h}}_j||}\right)}$$

where, as in prior work [96], we utilize a temperature $\tau = 0.07$.

C Training our Open-Source Toxicity Classifier

DeBERTa-based Contrastive Embedding Layer. Besides utilizing our augmented dataset of realistic adversarial perturbations, while training our model, we pre-train a contrastive layer to differentiate toxic and non-toxic texts. We later freeze this layer while training our full model to identify the toxicity of individual tweets.

To pre-train this layer for use in our model, we utilize contrastive learning to differentiate toxic and non-toxic texts. As in the original Civil Comments task, while training this layer we consider texts with labeled toxicity $t_i > 0.5$ score in the Civil Comments dataset as toxic and those with labeled toxicity $t_i < 0.5$ as nontoxic. We utilize this threshold for classifying a comment as toxic, given that this score (as described in the Civil Comments task) indicates that a majority of the Civil Comments annotators would have assigned a "toxic" attribute to this comment. For training, this is such that we embed each example $x_i = (text_i, t_i) \in D_{Civil_{aug}}$ (where $text_i$ is the text and t_i is whether the text is toxic or not) using a contextual word model by inputting $[CLS]text_i[SEP]$ and outputting the hidden vector $\mathbf{h}_i$ of the [CLS] token for each $text_i$ as its representation. Then, given a set of hidden vectors $\{\mathbf{h}_i\}_{i=0}^{N_b}$, where N_b is the size of the batch, we perform a contrastive learning step on that batch. This is such that for each Batch $\mathcal{B}$, for an *anchor* hidden embedding $\mathbf{h_i}$ within the batch, the set of hidden vectors $\mathbf{h_i}, \mathbf{h_j} \in \mathcal{B}$ vectors where $i \neq j$, we consider them a positive

pair if t_i, t_j are equivalent. Other pairs where $t_i \neq t_j$ are considered negative pairs. Within each batch $\mathcal{B}$, the contrastive loss is computed across all positive pairs in the batch such that:

$$L_{toxic} = -\frac{1}{N_b} \sum_{\mathbf{h}_i \in \mathcal{B}} l^c(\mathbf{h}_i)$$

$$l^c(\mathbf{h}_i) = log \frac{\sum_{j \in \mathcal{B} \setminus i} \mathbb{1}_{[t_i = t_j]} \exp(\frac{\mathbf{h}_i^\top \mathbf{h}_j}{\tau ||\mathbf{h}_i|| ||\mathbf{h}_j||})}{\sum_{j \in \mathcal{B} \setminus i} \exp(\frac{\mathbf{h}_i^\top \mathbf{h}_j}{\tau ||\mathbf{h}_i|| ||\mathbf{h}_j||})}$$

where, as in prior work [96], we utilize a temperature $\tau = 0.07$. Throughout training, we use a batch size of 64 and a learning rate of 1×10^{-5}, training for three epochs. After training this layer, we freeze it for use in the rest of our model. As seen in Figure 2, reducing the dimensionality of the outputted $h_{constrat}$ on the Civil Comments validation dataset using t-SNE [147], our contrastive embeddings are largely, though imperfectly, able to differentiate between non-toxic and toxic comments.

Full DeBERTa Toxicity Detection Model. Taking our pretrained-DeBERTa contrastive embedding layer and our augmented dataset $D_{Civil_{aug}}$, we finally train our full DeBERTa toxicity detection model (Figure 16). This model first computes the scaled dot product of a DeBERTa hidden representation of a text h_{text} and the $h_{contrast}$ output of our DeBERTa contrastive embedding layer. The intuition behind this approach is to enable our model to determine the extent of the toxicity features present within the original text.

$$r_{contrast} = \sum_i a_i h_{text}^{(i)},$$

$$a_i = \text{softmax}\left(\lambda h_{text}^{(i)} \cdot (W_{contrast} h_{contrast})\right)$$

where and $\lambda = 1/\sqrt{E}$, E = dimensionality of the embeddings, and $W_{contrast}$ is a learned parameter matrix. Finally, once $r_{contrast}$ is calculated, we concatenate it using a residual connection with the original h_{text}. We then feed the resulting representation into a feed-forward network with ReLU activation for determining the toxicity of the text as seen in Figure 16. We minimize mean squared error while training, utilizing the Civil Comments validation dataset to perform early stopping with a patience of 2. Throughout training, we use a batch size of 64 and a learning rate of 1×10^{-5}. We completed all training on an NVIDIA A6000 GPU.

D Pointwise Mutual Information

The pointwise mutual information PMI of a particular word $word_i$ in a cluster C_j is calculated as:

$$PMI(word_i, C_j) = log_2 \frac{P(word_i, C_j)}{P(word_i)P(C_j)}$$

where P is the probability of occurrence and a scaling parameter α is added to the counts of each word. This scaling parameter α prevents single-count or one-off words in each cluster from having the highest PMI values. Given the scale of our dataset and the number of clusters within our dataset, we determine that a baseline count of 1 (α =1) for each word in the full dictionary in each cluster led to the best results [146].

E DP-Means

DP-Means [86] is a nonparametric extension of the K-means algorithm that does not require the specification of the number of clusters *a priori*. Within DP-Means, when a given datapoint is

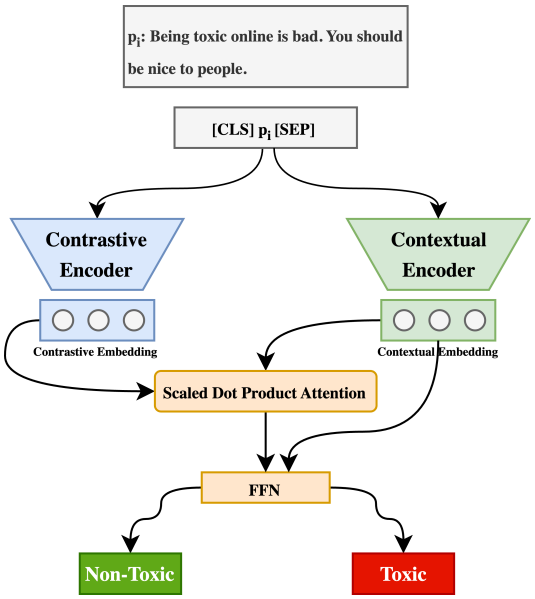


Fig. 16. Model to determine the toxicity of individual tweets— We utilize contrastive learning, scaled-dot-product attention, and the DeBERTa model to train a model to predict the toxicity of tweets in our dataset.

a chosen parameter λ away from the closest cluster, a new cluster is formed. Dinari et al. [33] parallelize this algorithm by *delaying cluster creation* until the end of the assignment step. Namely, instead of creating a new cluster each time a new datapoint is discovered, the algorithm determines which datapoint is furthest from the current set of clusters and then creates a new cluster with that datapoint. By delaying cluster creation, the DP-Means algorithm can be trivially parallelized. Furthermore, by delaying cluster creation, this version of DP-Means avoids over-clustering the data (*i.e.,* only the most disparate data points create new clusters) [33].

F GAM Fit of User Level Features and Perspective Toxicity

Train R^2 0.266, Validation R^2: 0.270			
Dependent Variable	Pearson Corr. ρ	Kendall's τ	Permut Import.
Verified Status	—	-0.233	0.031
Years Active on Twitter	-0.220	-0.155	0.022
Log # Followers	-0.229	-0.137	0.231
Log # Followed	-0.197	-0.128	—
Log # Tweets in 2022	0.182	0.173	0.094
Toxicity of Mentioned Users	0.366	0.347	0.409
Partisanship	0.075	0.079	—
σ(Mentioned Users Partisanship)	0.331	0.294	0.149
μ|User Partisanship- Mentioned Partisanship|	0.272	0.241	0.015
μ(Mentioned Partisanship)	0.139	0.114	0.048

Table 5. Pearson correlation ρ, Kendall's τ, and permutation importance of dependent variables and users' toxicities. As seen in the above table, a user's interaction with a wide political variety of users and interacting with other users with higher toxicity correlates with a given user's own toxicity.

G Most Toxic Topics

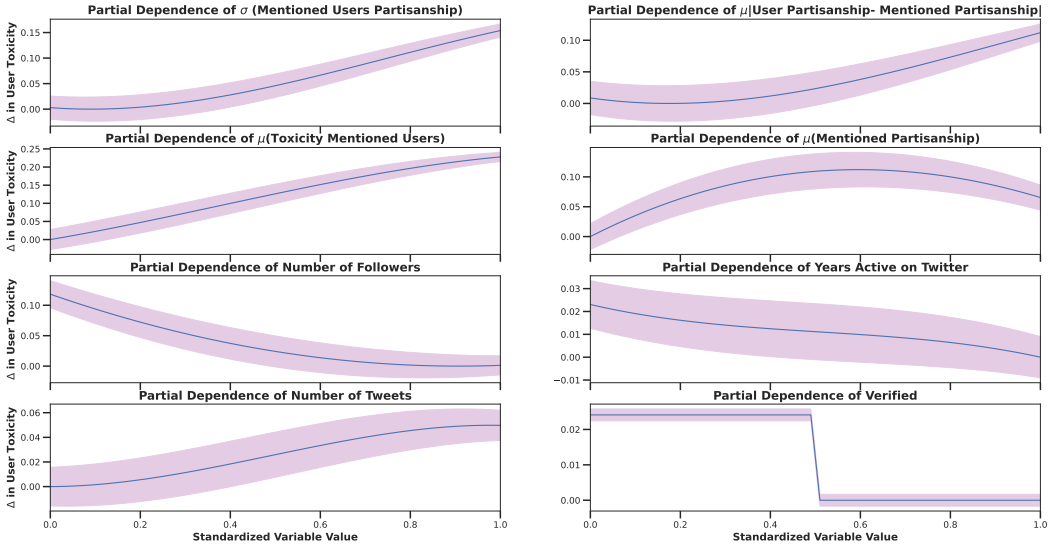


Fig. 17. Partial dependencies with 95% Normal confidence intervals between fitted standardized dependent variables and user Perspective API toxicity.

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	fuck, shit…, shit, shitttttt, extremely	52	52 (100%)	0.923	That's all folks. Fuck this shit.	-0.169	-0.167	0.737
2	idiot, blithering, complete, total, he	3121	3,123 (99.94%)	0.915	Not idiots. Deliberate enablers of fascism.	0.152	0.128	0.970
3	fuck, you, him, though, that's	340	336 (98.82%)	0.902	Fuck this and fuck him.	-0.027	-0.117	0.836
4	piece, load, shit, hahaha, you	756	775 (97.55%)	0.895	Tell me you are a piece of shit without telling me.	0.011	0.007	0.957
5	volume, youtube, chop, stupid, that	435	438 (99.32%)	0.880	Nothing stupid about that!	0.119	0.104	0.942

Table 6. Top toxic topics—by average toxic value—in our dataset.

H Linear Fit of Topic Level Features against Perspective Toxicity

Train R^2 0.454, Validation R^2: 0.463			
Dependent Variable	Pearson Corr. ρ	Kendall's τ	Permut Import.
Number of Users	-0.268	-0.132	0.520
μ(Years Active on Twitter)	-0.233	-0.192	0.010
Percentage Verified	0.273	0.247	0.014
σ(User Partisanship)	-0.097	-0.012	0.036
μ(User Partisanship)	-0.014	0.011	0.013
μ(User Toxicity)	0.637	0.502	0.398

Table 7. Pearson correlation ρ, Kendall's τ, and permutation importance of dependent variables and clusters' toxicities.

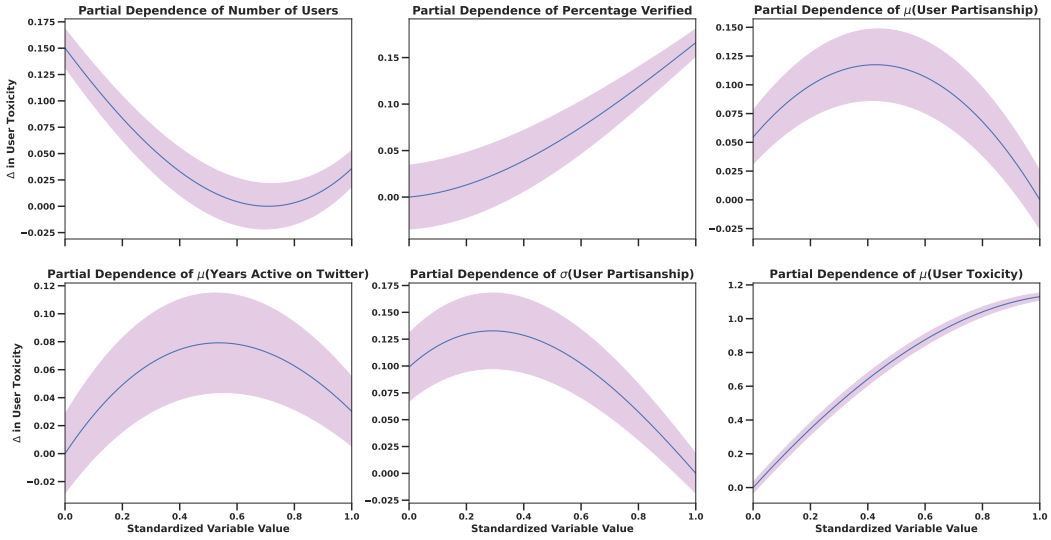


Fig. 18. Partial dependencies with 95% Normal confidence intervals between fitted standardized dependent variables and cluster Perspective API toxicity.

Received October 2024; revised April 2025; accepted August 2025