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 hyperpartisan users post pernicious content, slanted to their political views, inciting contentious and toxic conversations. However, what factors actually contribute to increased online toxicity and negative interactions? In this work, we explore the role that political ideology plays in contributing to toxicity both on an individual user level and a topic level on Twitter. To do this, we train and open-source a DeBERTa-based toxicity detector with a contrastive objective that outperforms the Google Jigsaw Persective Toxicity detector on the Civil Comments test dataset. Then, after collecting 187 million tweets from 55,415 Twitter users, we determine how several account-level characteristics, including political ideology and account age, predict how often each user posts toxic content. Running a linear regression, we find that the diversity of views and the toxicity of the other accounts with which that user engages has a more marked effect on their own toxicity. Namely, toxic comments are 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 behavior across 6,592 individual topics, with conversations on each topic becoming more toxic as a wider diversity of users become involved.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing*; **Empirical studies in collaborative and social computing**; • **Information systems** → **Web Mining**; • **Networks** → *Online social networks*;

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

1 INTRODUCTION

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

Over the past decade, political polarization within the United States has increased substantially with many blaming social media for the increase in division [10, 14, 27–29, 39]. Social media, several argue, creates toxic political echo chambers where users become more politically polarized and where users' biases are reinforced [80, 90]. In several document cases, this perceived high degree of political polarization and toxicity negatively has heavily impacted platforms, online communities, and users, sometimes leading to users leaving platforms altogether [24]. While many studies have investigated the role that toxicity and political polarization have had on the health of online communities [31, 62, 73, 83–85], there has been little work that investigates the role of toxicity, political ideology, and affective polarization (*i.e.*, the tendency to negative to those with different political views and positive to those with similar political views) at the individual user and topic-level. To fully understand the intertwined relationship between toxicity, political ideology, and polarization, at the user and topic-level in this work, we ask the following research questions:

(1) *What are the most prominent factors that predict the tendency of politically engaged users to post toxic content on Twitter?*
(2) *What topics and conversations on Twitter engendered the most toxicity in 2022? How did the characteristics of the users engaged in these conversations affect the toxicity of these topics?*

Authors' addresses: Hans W. A. Hanley, hhanley@stanford.edu, Stanford University, 450 Serra Mall, Stanford, California, USA, 94305; Zakir Durumeric, zakird@stanford.edu, Stanford University, 450 Serra Mall, Stanford, California, USA, 94305.

To address these questions, we collect 187 million tweets from 54,515 accounts throughout 2022. From these tweets, we measure the number of toxic tweets and toxicity of each user by designing and deploying our own DeBERTa [38] toxicity detection model, finding that it outperforms Google Jigsaw API [1], the gold-standard out-of-box classifier for identifying uncivil and toxic language (*e.g.*, insults, sexual harassment, and threats of violence [82]). Then calculating each user's approximate political orientation using Correspondence Analysis [8] 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. To begin, we first determine, using a linear regression model, some of the most significant features that predict the toxicity of the content posted by individual Twitter accounts. We find that the most important feature that predicting an individual account's toxicity is the toxicity of the other accounts with which the user interacts (accounting for 13.96% of the variation). 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 own messages, the more that a given user interacts with users of different political orientations, the more toxic their own content tends to be.

RQ2: Toxicity and Political Polarization Within Toxic and Malign 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 6.6M English-language toxic tweets in our dataset, we perform topic analysis using a fine-tuned version of the large language model MPNet and the DP-Means clustering algorithm [35]. Examining these topic clusters, we find, in aggregate, that the political orientation of the users tweeting about each topic does not have a large effect on the overall each topic's overall toxicity; rather we find that the effect of the political orientation of the users tweeting about particular topics varied widely. Examining factors that predict each topic cluster's overall toxicity, we find that the average toxicity of the users tweeting about that topic and the variance in the political views of those same users positively correlated with each topic's toxicity. Namely, we find at the topic-level (as on a user-level), a 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. Our work, one of the first to perform this analysis on a large-scale dataset of politically engaged users, 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.

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 [40, 52, 73, 92].

Political Ideology: As in Barbera *et al.* [7] and other works [71, 72], we define US political ideology 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 [63], and this assumption is fairly common in the literature.

Affective Polarization: Affective polarization is the tendency of individuals to distrust and be negative to those of different political beliefs while being positive towards people of similar political views [23].

2.2 Twitter

Twitter is a microblogging website where users can post messages known as Tweets. These tweets can consist of messages with at most 280 characters. Tweets themselves, while often just text, can also include hyperlinks, videos, and other types of media [46]. Unless made private, tweets are publicly displayed on the Twitter platform, allowing anyone to see or reply to the message [48]. As of late 2022, Twitter had approximately 238 million active daily users [20]. Many Twitter users get their daily news from the Twitter platform [4, 11, 81]. 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 [19].

2.3 Political Ideology and Polarization Online

Various works have explored the role that individual users' political orientations in interactions online. People, on the Internet and in their everyday interactions, tend to associate and be friends with like-minded individuals and Twitter is no exception [6, 8, 34, 41, 47, 64]. Several works have found that social media, exacerbates this human tendency, by creating political echo-chambers [78], where users' biases are reconfirmed and reinforced [5, 9, 15, 17]. Cass Sunstein, Garett *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 [26, 65, 80]. While the vast majority of Twitter users do not engage in political discussions, those that do, are often highly politically polarized, rarely following or engaging with different political beliefs [90].

In addition to polarization being amplified by social media, other works have found this increased polarization can increase misinformation and toxic behavior [5]. Rains *et al.* [66], for instance, find that high polarization is a major factor in engendering online incivility and toxicity. Imhoff *et al.* [43], 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 [18, 53, 61, 82, 91]. 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 [50, 54, 74].

Several works measure online toxicity using Perspective API [1]. Saveski *et al.* [73], 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.* [33], utilize Perspective to identify opportunities for proactive interventions on Reddit before large escalations. Kumar *et al.* [53] 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.

2.5 Present Work

Several works, close to our study, have attempted to understand how political polarization and online toxicity interact online in particular political environments [16, 84]. For example, Chen *et al.* [14] utilize network analysis to find that misleading online videos lead to increased online incivility. Conversely, Rajadesingan *et al.* [67], find that political discussions in non-overtly political subreddits often lead to less toxic conversational outcomes. Most similar to our work, De Francisci Morales *et al.* [21] find, that the interaction of individuals of different political orientations increased negative conversational outcomes. In this work, however, rather than examining political polarization within a particular community or across one particular topic, we instead seek to understand across thousands of politically engaged users across the political spectrum, what are the most prominent characteristics that correspond with increased toxicity. By then extending this to a topic-level analysis, we examine the most how these account features correspond to the toxicity of conversations online 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 Political Ideology

To approximate individual Twitter users' political ideology, we rely on the Correspondence Analysis (CA) and proposed by Barberá *et al.* [8]. Correspondence analysis (CA), similar to principal component analysis, is a technique for categorical data that extracts discriminating and representative features from a given matrix [30]. 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 can determine a discriminating latent space among these users based on their following behaviors. By carefully choosing our set of "followed" users (columns of the matrix) as a set of key political figures, this latent space can be used to represent a dimension of "political ideology." 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, and vice versa [8, 59]. We note that with the CA technique, by later extending the set of the key followed accounts, this approach can be used to approximate the political ideology 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 political 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 55,415 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 Democratic political leaders as well as our set of 55,415 accounts. In this setup, the more positive a

[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

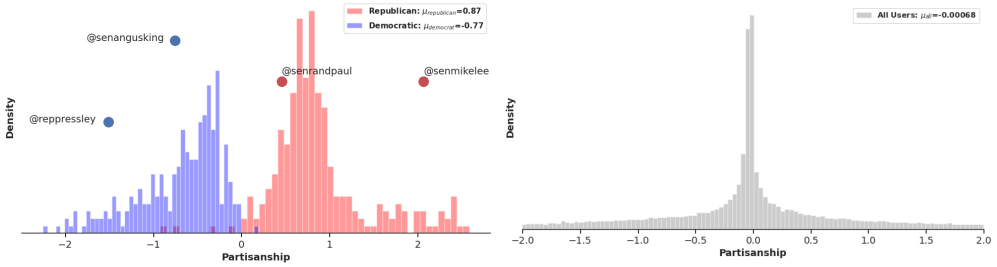


Fig. 1. Estimated Political Orientation of Political Leaders and All Users Using CA– We differentiate users' political leanings based on who they follow on Twitter.

user's ideology, the more right-leaning; conversely the more negative a user's ideology, the more left-leaning.

3.2 Collecting Tweets

Our Twitter dataset consists of tweets from 55,415 Twitter users from throughout 2022, altogether 187,628,895 tweets. Each user was selected randomly for a set the users that followed our key political figures. See Appendix A for details. For each user, using the Twitter API, we gather all available tweets for the user from 2022. While we acknowledge several of our users' tweets might have been deleted or taken down by moderators before we scraped them, this dataset, consisting of over 187,628,892 tweets, with an average of 3385.8 (median 1018.0) tweets per individual is largely comprehensive of each user's tweet behavior. Using the `whatlango`[3] library, we find that 62.2% (116.5M) of our Tweets were in English, 4.87% were in French, 2.39% were in Spanish, and the rest were in an assortment of different languages. We note that for much of our analysis throughout this paper, we rely on **only** the set of tweets that are in English, discarding the remaining tweets.

3.3 Identifying the Toxicity of Tweets

3.3.1 Designing an Open-Source Toxicity Classifier. We design and open-source[4] a contrastive DeBERTa-based [38] model to determine the toxicity of tweets, later benchmarking our approach on two public datasets and against the Perspective Toxicity API [1], the gold standard of toxicity detection [1, 53, 68]. We note that throughout our work, we reproduce several results using the Perspective Toxicity classifier and present them in the Appendix after obtaining similar results.

To train our new model we rely on the Civil Comments dataset[5] that was also utilized to train and validate the Perspective API. In addition to utilizing this dataset to augment our trained model, we take two main approaches: (1) data augmentation through realistic adversarial perturbations of the original Civil Comments dataset [55], and (2) the inclusion of a contrastive learning embedding layer to help better differentiate toxic and non-toxic texts.

Realistic Adversarial Perturbations of the Civil Comments Training Dataset. To train our model, we rely on the Civil Comments training dataset which consists of 1,804,874 comments that were each individually graded by up to 10 human raters for their toxicity. Each comment, depending on the percentage of 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. Our training dataset is thus D_{Civil} =

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

$\{x_i = (text_i, t_i)\}_{i=1}^{N}$, where $text_i$ a text, and t_i is the toxicity of the text. While the Civil Comments training dataset is fairly large, we note that it is heavily skewed with 1,268,269 of the comments having a toxicity score of 0. To ensure that our training dataset has a wider set of examples of comments with above zero estimated toxicity, we augment the Civil Comments training dataset using realistic adversarial perturbations [55].

Utilizing the ANTHRO dataset provided by Le *et al.* [55], for every comment with above zero toxicity within the Civil Comments dataset, we leverage the set of common human-written perturbations to augment our Civil Comments dataset. This 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, we extract a set of random perturbations of each noun and adjective within the comment, perturbing the overall comment nine times with different combinations of the perturbed nouns and adjectives. This enables us to extend the set of non-zero comments to a total of 5,366,050 comments (6,634,319 in the full augmented dataset). We utilize this dataset when training our DeBERTa-based [38] model to determine the toxicity of tweets. 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.

A 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. 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:

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

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

where, as in prior work [57], 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 [87], our contrastive embeddings are largely though imperfectly, able to differentiate between non-toxic and toxic comments.

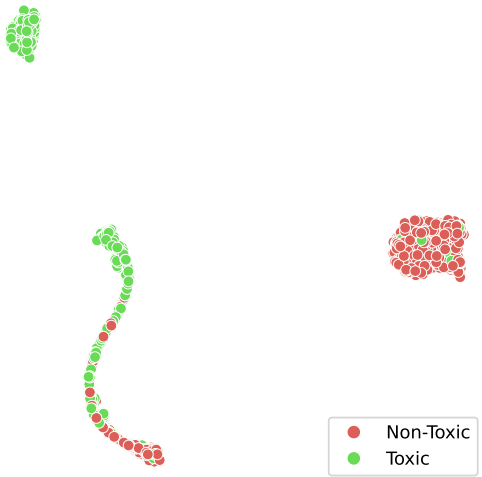


Fig. 2. **t-SNE of Civil Comments Validation Dataset** – As we train the DeBERTA-based contrastive embedding layer of our model on our augmented Civil Commonents training set, our model is able to differentiate non-toxic (*i.e.,* toxicity $t_i < 0.5$) and (*i.e.,* toxic $t_i > 0.5$) comments. However,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

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 3. 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 = dimentionality of the 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 3. 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 a Nvidia A6000 GPU.

3.3.2 Benchmarking our Toxicity Classifier. Upon training our DeBERTa-constrative toxicity model, we benchmark it against the Perspective Toxicity API [1] as well as a vanilla finetuned DeBERTa model with a classification head (a two-layer MLP with ReLU activation). To benchmark our toxicity model, we utilize the validation and test dataset of the Civil Comments dataset provided by Google Jigsaw[1] as well as a separate toxicity dataset provided by Kumar*et al.* [53]. Kumar *et al.*'s [53] 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 annotaros in the Civil Comments dataset). For our F_1 score calculations, as in Kumar *et al.* [53] and in the Civil Comments dataset, we consider a comment to be toxic if its toxicity $t_i > 0.5$.

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

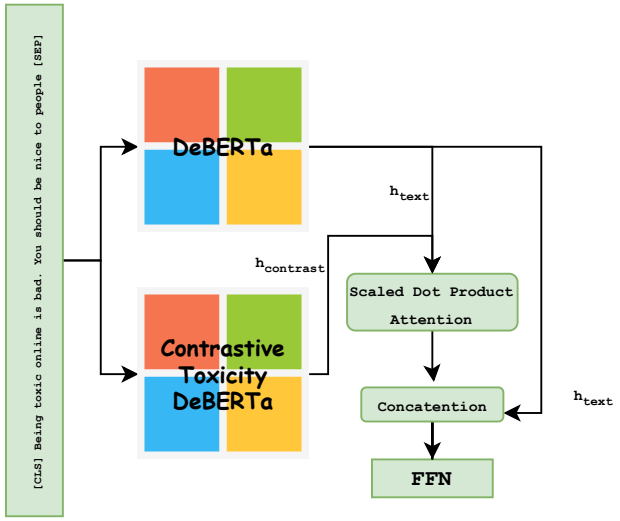


Fig. 3. Model to determine the toxicity of individual tweets— We utilize contrastive learning, scaled-dot-product attention, and the DeBERTa model to train a model to predict the toxicity of tweets in our dataset. Our fully trained model achieves a 0.818 Pearson correlation with the toxicity scores in the Civil Comments test dataset.

	CC Validation			CC Test			Kumar *et al.*		
Model	MAE	Corr.	Macro-F_1	MAE	Corr.	Macro-F_1	MAE	Corr.	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**
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

test dataset. In addition, while obtaining a slightly lower correlation, our model on this separate dataset achieves a lower mean absolute error and a higher F_1 score. As such for the rest of this work, when determining the toxicity of tweets, we utilize our contrastive DeBERTa model. We note that, as in other works [36, 68], 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 DPMeans

To later understand how particular types of users interact with different topics composed of toxic tweets (as labeled by our model), we perform topic analysis on these messages. As found by Grootendorst *et al.* [32, 35], 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[6] fine-tuned on semantic search and a parallelizable minibatch version of the DP-Means algorithm.[7]

3.4.1 MPNet. To compare two tweets' semantic content for later clustering, we rely on a version of the MPNet [76] 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[8] that we utilize was fine-tuned on similar social media data (*e.g.*, Reddit comments and Quora Answers) allowing us to apply this model to

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

0.55 similarity

Tweet 1: So this speakership battle is interesting and all, but we still have a massive crisis on our southern border that the Biden administration is doing nothing about. From:@benshapiro
Tweet 2: Amid the worst southern border crisis on record, Democrat Rep. Adam Schiff insists the border "has been a very strong priority" for Democrats and praises Biden for "working in good faith." From: @RNCResearch

-0.018 similarity

Tweet 1:I've just spent the night listening to Pokémon Ruby and Sapphire music, reminiscing over my wonderful childhood, only to just find out this game came out 20 YEARS AGO. I hate getting old. From: @TheAn1meMan
Tweet 2: POV: You're Nancy Pelosi talking to Kevin McCarthy From @santiagomayer_

Fig. 4. Examples of Tweet pairs at different similarities (0.55 left and -0.018 right).

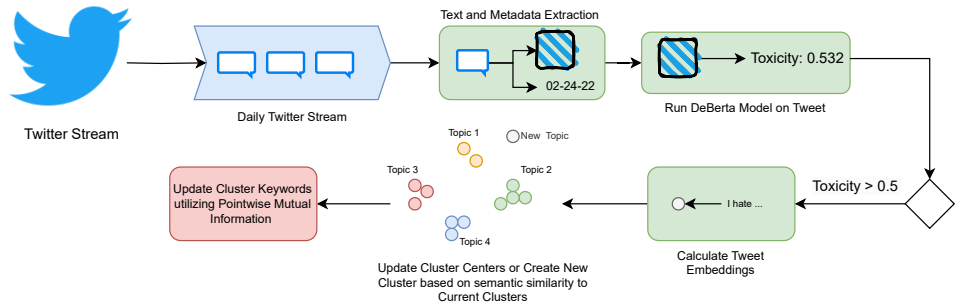


Fig. 5. Topic analysis of Toxic Tweets—We determine the toxicity, embed, and cluster toxic tweets to identify the most polarized and toxic conversations on Twitter throughout 2022. We note that for this approach, we limit our analysis to English tweets. We utilize the `whatlango` Go library to determine the langue of tweets.

our set of tweets. As a reference, we provide two example tweet pairs with similarities at 0.55 and -0.18 in Figure 4. 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.

3.4.2 DPMeans. DP-Means [22] is a non-parametric 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. This characteristic of DP-Mans 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 non-parametric 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 C.

3.4.3 Human Understandable Clusters: Keyword Extraction and Representative Tweets. We note after clustering tweets, to make clusters 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 [32]. Second, we determine the most distinctive keywords of each cluster using pointwise mutual information [12] (detailed in Appendix B). In this way, after clustering our set of tweets, we can later extract the semantic meaning of the various clusters outputted.

3.4.4 Topic Analysis Pipeline. Having outlined the constituent elements of our topic analysis algorithm, we now go over the full topic analysis pipeline (Figure 5): Throughout 2022, as we gathered the tweets of our set of 55,415 Twitter users, using our DeBETa-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 `whatlango`, using MPNet, we subsequently map these tweets to a shared embedding space. Finally, we continuously cluster and identify topics amongst these toxic tweets using the DP-Means algorithm. As recommended

by Hanley *et al.* we utilize a λ of 0.60 for our clusters (precision near 0.989 for MPNet [32, 37]). 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 6,694,756 English-language toxic tweets from our set of 55,415 Twitter users, we identified 6,592 cluster centers with at least 5 tweets.

3.5 Ethical Considerations

Within this work, we largely focus on identifying large-scale trends in how different Twitter 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 lastly note that our Twitter data was largely collected prior to Elon Musk's private acquisition of Twitter on October 27, 2022, and all of our data was collected prior to the later restrictions placed on the collection of tweets on June 30, 2023.[9]

4 RQ1: USER-LEVEL FACTORS IN TOXICITY ON TWITTER

Having provided background on our methodology and dataset, in this section, we discuss several of the user-level factors that coincide with and contribute to the 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 contributing to or mitigating how toxic particular users are in their interactions on Twitter:

(1) The verified status of the account
(2) The amount of time in years that the account has been active on Twitter
(3) The log of the number of followers that the account has on Twitter
(4) The log of the number of other Twitter users that the account follows
(5) The absolute value of the account's political ideology as determined by our Correspondence analysis
(6) A binary value of whether the account leans politically left or right
(7) The estimated average toxicity of all users the account mentioned/@ed on Twitter (*i.e.*, accounts that the user has interacted with)
(8) The standard deviation of the polarization of the accounts that the account mentioned (*i.e.*, the range of political views the user interacts with
(9) The average difference in the political ideology of users the given account has mentioned and the account's own political ideology
(10) The average absolute value of the political orientation of the mentioned accounts

We fit these 10 covariates against each of our account's average toxicity scores. We note that throughout this analysis, amongst our 55,415 accounts, we only include accounts that mentioned or interacted with another one of our 55,415 accounts, altogether 43,381 users.

To understand how these factors interact with and contribute to toxicity on Twitter, we fit a linear regression on the average toxicity score of users against our set of user characteristics (Table 2). In addition to fitting this linear regression, we further determine the estimated amount of variance explained by each variable. As seen in Table 2, we do indeed observe that each of the user characteristics that we consider to varying degrees *does* indeed have observed correlational effect on how toxic users' tweets tend to be. We consider each of these effects below. We lastly note that in order to ensure the robustness of our approach, we separately perform the same analysis

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

Adjusted R-squared: 0.259	Coefficient	Std. Error	Adj. Sum Sq.
Intercept	0.0481	0.002	1.599
Verified Status	-0.0100	0.001	0.241
Years Active on Twitter	-0.0009	-0.00007	0.691
Log # Followers	-0.0108	-0.001	1.368
Log # Followed	-0.0061	-0.001	0.269
Log # Tweets in 2022	0.0063	0.001	0.614
Toxicity of Users in Mentions	**0.7744**	0.012	**11.437**
Abs Political Ideology	-0.0023	0.001	0.068
Left/Right	-0.0013	0.001	0.017
Std. of Mentioned Users Political Ideology	0.0191	0.001	1.014
Avg Abs(User Ideology - Mentioned Users Political Ideology)	0.0307	0.001	3.203
Abs Avg of Mentioned Users Political Ideology	-0.0156	0.001	0.695

Table 2. Linear fit of several user-level factors on average toxicity of Twitter users. All coefficients had a p-value ≈ 0 using t-tests. We perform analysis of variance (ANOVA) and t-tests to further determine significance and estimate the variation that can be attributed to various factors. (All coefficients are indeed significant and contribute to explaining toxicity on Twitter).

utilizing the toxticity scores output by the Perspective API, obtaining similar results. We present those results in Appendix D.

4.2 Baseline Account Characteristics

We first provide an overview of how several baseline account characteristics contribute to the toxicity of each user.

Verified Status. As seen in Table 2, as also found by Aleksandric *et al.* [2], 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 see that whether a user is verified or not explains 0.29% of the variation in toxicity of the content of individual users' tweets. We note that we collected users verification status prior to the implementation of Twitter Blue (users could pay $8 USD to become verified) in November 2022 [25].

Years Active on Twitter. As users stay on Twitter, we observe that they are slightly less likely to be toxic. As argued by Rajadesingan *et al.* [68] 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 see that the number of years that a particular user has been on Twitter explains 0.84% of variability in user toxicity on the platform.

Number of Followers. Like verified status, and as argued by Marwick *et al.* [58], 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. We see the same: users become more popular and have more followers, and they are less likely to post toxic tweets on their profiles. This variable explains 1.67% of the variability of toxicity of users in posting toxic content online.

Number of Users Followed. As found in a recent analysis of Twitter by Saveski *et al.* [73], users with fewer social connections are more likely to be toxic. We observe a similar phenomenon: accounts that follow a smaller number of users are more toxic than those that follow more. Altogether, the number of accounts followed accounts for 0.32% of the variability in user toxicity on Twitter.

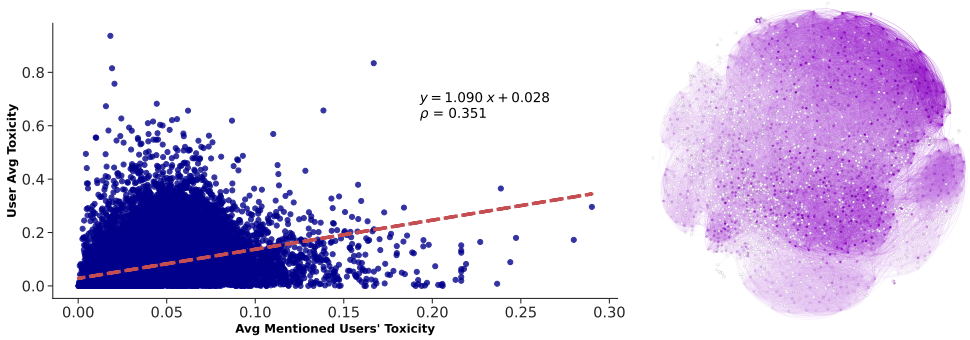


Fig. 6. The more toxic the users mentioned by a given user, on average, the more toxic the content of that particular user. Within the mention graph of user interactions, toxicity has an assortativity coefficient of 0.0661, suggesting that, to some degree, users who post toxic content tend to mention and interact with other users who post toxic content.

Number of Tweets. Many accounts in our Twitter dataset post several times a day, with the median account posting 1,018 times throughout 2022, and one account posting 754,905 times. We observe that as Twitter users post more, their average toxicity increases. This finding reinforces past work that suggests that accounts that post excessively are more likely to be toxic [69]. Altogether, the number of tweets posted accounts for 0.75% of the variability in user toxicity.

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 that post toxic content, they themselves are more likely to be toxic. Altogether, this one covariate accounts for 13.96% of the variability of toxicity (Figure 6). 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 [68, 73]. Aleksandric *et al.* [3] for example, find that merely observing toxic online behaviors has the effect of increasing toxic interactions. Creating a mention (@) graph among our 43,381 users, we indeed find some degree of assortativity based on toxicity (0.0661), with more toxic users more likely to interact with each other than with non-toxic users.

Political Ideology. We *do not* see that the most politically ideological Twitter users in our dataset are the ones who tweet the most toxic content. Rather, political ideology appears to account for very little of the variation in user toxicity as seen in some prior work [36]. While there is a slight trend for users with more politically insular views to be less toxic, we find that overall this explains very little of individual users' toxicity. Altogether, the place of particular users on the political spectrum, in absolute terms, accounts for 0.02% of variability in toxicity.

Left vs Right. Overall, we find that right-leaning users are slightly more toxic than left-leaning users on Twitter. As seen in Figure 7, using our toxicity classifier, we find that right-leaning users have slightly higher toxicity than liberal-leaning users. We observe that right-leaning users are toxic at a rate 1.18 times higher than left-leaning users (Cohen's D = 0.160, $p \approx 0$[10]). We thus observe

[10]p-value was calculated using Mann Whitney U-test

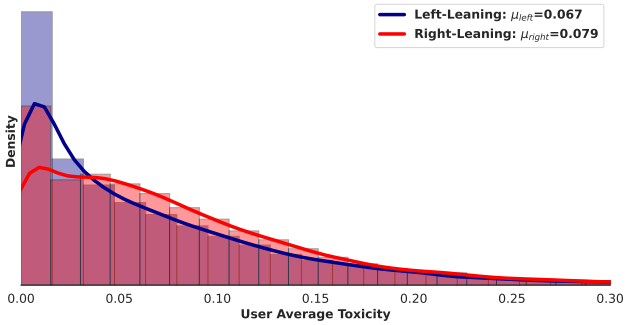


Fig. 7. Toxicity of Right-leaning and Left-leaning Twitter Users—Right-leaning users post toxic tweets slightly more frequently than left-leaning users.

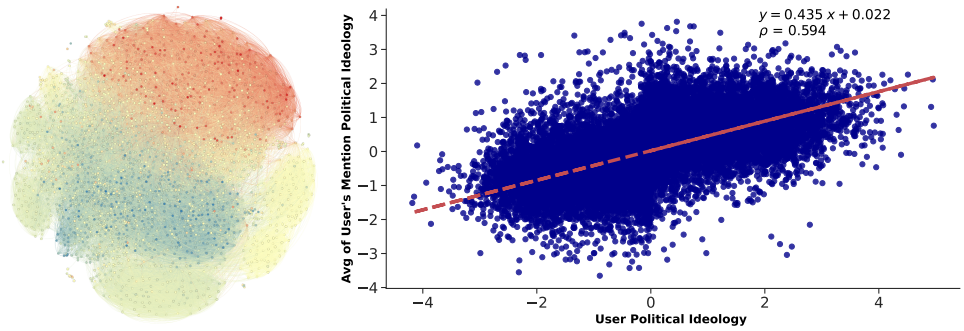


Fig. 8. Within the mention graph of user interactions (red/right-leaning and blue/left-leaning), political ideology has an assortativity coefficient of 0.258, suggesting that conservative users mention and interact more with right-leaning users while liberal users interact more with and mention other left-leaning users. Similarly, graphing the average of each user's mention's political ideology against their own political ideology, we find significant assortativity (Pearson correlation $\rho = 0.594$)

a small, but measurable effect based on the political leaning of particular users. This covariate, however, accounts for only 0.08% of the variability of user toxicity. We thus again find users' being right-leaning or left-leaning largely does **not** explain much of the toxicity on Twitter.

Referencing the Political Extremes. Altogether, the average absolute value of the political orientation of a user's mentions explains 0.84% of the variability in a given user's average toxicity. We observe, however, 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 less toxic content on Twitter. This may do with 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 43,381 users, we find some degree of assortativity (0.258), thus finding that users, on the whole, tend to interact with other users of similar political views (Figure 8). Graphing the average political ideology of a user's mention against their own political ideology we further observe a high assortativity (Pearson correlation of $\rho = 0.594$).

The Political Diversity of Mentions. From Table 2, we find that as users mention (@) a wider political diversity of users, the more toxic their own tweets tend to become (Figure 9). The political diversity of users mentioned by any given user accounts for 1.23% of the variability in user toxicity.

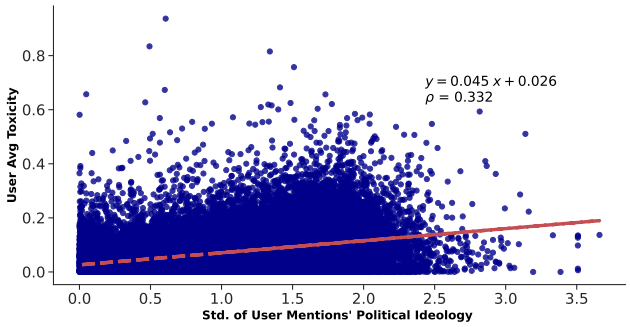


Fig. 9. As users mention a wider range of users along the political spectrum they are more likely to tweet toxic messages.

We similarly see that as users mention and interact with users that are more different from themselves, they more likely that they are to be toxic. Simply put, as a user's typical interaction with other users moves to involve other users of different political orientations, the more toxic content that user tweets. This particular covariate accounts for 3.91% of the variability in user toxicity. This appears to indicate some degree of affective polarization with our data; this covariate accounts for the most variation in toxicity besides the toxicity of other users (Table 2). Together these two variables suggest that toxicity tends to increase not only when users interact with other users of wide-berth political ideologies, but also when users interact with users politically different from themselves.

4.4 Summary

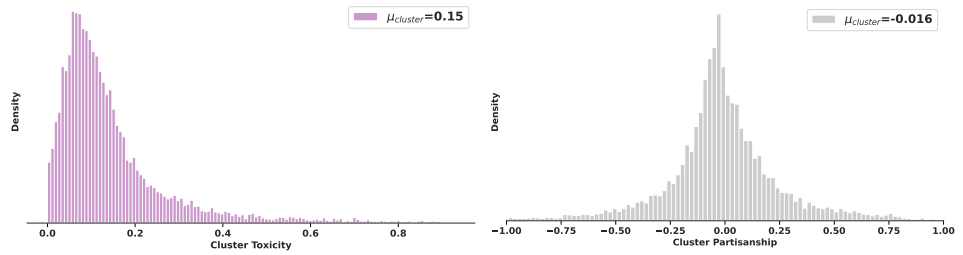
In this section, using a linear regression model, we explored the role that several user-level characteristics have on the rate of user toxicity on Twitter. We find, most importantly, that users who interact and mention other users who regularly post toxic content are more likely to be toxic themselves. Similarly, we found 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 D.

5 FACTORS AND CHANGES IN POLARIZED AND TOXIC TOPICS ON TWITTER

Having investigated the role that various user characteristics have in user toxicity on Twitter, we now explore how different characteristics affect different negative and toxic conversations and topics on Twitter. Specifically, how does the toxicity of topics on Twitter change based on the makeup of the user participating in these conversations? Within this section, 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 6,694,756 toxic tweets from our set of 55,415 Twitter users, we identified 6,592 clusters with at least 5 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



(a) (b)

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

toxic tweet clusters, we now give a brief overview of how we estimate the overall toxicity and partisanship of each particular topic after identifying their corresponding cluster of toxic tweets.

Estimating the toxicity of topics. To estimate the toxicity of particular topics, we determine the average toxicity of all tweets present within that given cluster. In addition to this metric, we further determine the *percentage* of toxic tweets within our *entire* English-language dataset that conforms 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.

To assign non-toxic tweets to our set of toxic tweet centers, we utilize the approach laid out in prior work [35, 37] 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.* [37], given our particular version of MPNet, we again utilize a cluster threshold of 0.60 for assigning a given non-toxic tweet to a given cluster (precision near 0.989 for MPNet [37]). We plot the distribution of estimated topic toxicity in Figure 10a. We utilize this approach, rather than clustering all 116 million English tweets given the size of our dataset, and because, for this work, we largely are only concerned with topics that have some level of toxicity.

Estimating 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/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 political ideology [Section 3.1]) that tweeted about that topic, we thus estimate the political-ideological composition of each topic. We plot the distribution of the political ideology of our set of clusters in Figure 10b.

5.2 The Most Toxic Topics of 2022

Before examining the different factors that contribute to the toxicity of individual topics, we start this section by providing an overview of the topics with the most toxic tweets in 2022 (Table 3). We further list the set of topics with the highest average toxicity in Appendix E (we do not discuss these topics here as most of them 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 [60], namely Russia's invasion of Ukraine (Topic 2; 1.57M tweets), Joe Biden's administration (Topic 3; 797K tweets), and the raid of former President Donald Trump's home for classified documents by the US Federal Bureau of Investigation (FBI) (Topic 7; 776K tweets).

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	gop, party, democrat, republican, dems	897,606	97,555 (10.87%)	0.163	@REDACTED Filthy dirty gross disgusting Democrats that's who	-0.0522	0.0727	1.169
2	putin, ukraine, russia, putin, soldier, kyiv	1,779,569	93,827 (5.27%)	0.086	It's interesting that so many of those pushing so hard against any kind of military assistance to the freedom fighters of Ukraine are the same ones who cosplay as militants; walk around acting tough open carrying weapons at political demonstrations. What a bunch of ninnies.	-0.119	-0.0003	0.8226
3	joe, biden, administration, oil, senile, president	797,926	73,591 (9.22%)	0.136	@REDACTED @REDACTED I'm not saying to ignore Trump. I'm saying that he isn't the story right now. He's a side note to that story. Biden's in the WH. This is happening on Biden's watch. Call Trump the moron he is, then move on to the actual topic at hand.s	0.5491	0.300	1.010
4	potus, president, donald, trump, impeached	401,943	61,120 (15.21%)	0.222	@POTUS You have destroyed this country, you should resign and hang your head in shame	0.238	0.120	1.076
5	black, racist, supremacist, white, kkk	267,925	55,550 (20.73%)	0.326	democrats are racist	0.114	0.113	0.995
6	troll, bezos, biden, fuck, pedo	184,701	51,232 (27.74%)	0.338	@REDACTED @REDACTED Your pathetic way of thinking is the problem!!? #PedoHitler	0.146	0.0775	0.827
7	FBI, document, classified, committee, doj	776,153	48,290 (6.22%)	0.091	You mean they are finally going after Obama and the Clintons? About time they went after those crooks	0.0646	0.0797	1.113
8	tory, labour, snp, brexit, scotland	453,121	46,046 (10.16%)	0.131	@REDACTED Spot on! It's getting ridiculous with the virtue signalling shite!..	-0.011	0.0250	0.386
9	CNN, fox, msnbc, news, network	485,724	41,284 (8.501%)	0.139	@FoxNews And? You're all disgusting and Jesus hates you	0.0637	0.110	1.162
10	justin, trudeau, canada, ctvnews, ndp	457,283	40,992 (8.96%)	0.142	Mirror Mirror on the wall, Trudeau is a gaslighting liar after all!	0.223	0.371	0.582

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

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	demonrat, demoncrats, demoncraps, sado-masochistic, ultramaga	1056	167 (15.81%)	0.289	YelF DemonKKKrats COULDN'T LIE.. what would they have to talk about ? If they COULDN'T cause FRICTION; FIGHTING among people.. what would they have for policy ? DemonKKKrats are evil sick people.	0.913	0.810	0.700
2	gutfeld, colbert, over-rated, idiot,jackass	18,780	171 (0.91%)	0.077	Greg Gutfeld Lets Loose On 'Jackass' Biden After SOTU: 'You Are An Idiot' To Believe He Supports Cops	0.901	0.606	0.888
3	putz, ass, scrunt, dog-fight ,fuckers	5996	136 (2.27%)	0.040	@REDACTED Because they suck	0.880	0.857	0.740
4	psyop, pepe, spook, muh, cough	6,970	258 (3.70%)	0.100	@REDACTED Turn his butt into the authorities…	0.879	0.942	0.766
5	rino, worthless, ultra-maga, establishment, backstabbing	21,114	2039 (9.66%)	0.167	THESE TWO FEDGOV SHITSTAINS WILL ATTEMPT TO TURN HIM INTO ANOTHER FKN rino, a demoncrap in pub clothing!	0.860	0.734	0.869

Table 4. Top toxic topics with at least 100 tweets—by right-leaning tilt—in our dataset.

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 (0.0157 rightward tilt). Examining these tweets, we find right-leaning users, as seen in the example tweet (Table 3), when tweeting about the war, excoriated or derided the Ukrainian government or military, which was picked up as toxic by our contrastive-DeBERTa model. In contrast, considering all users who tweeted about the war, we find that they tended to lean leftward (-0.119 leftward tilt).

In contrast, looking at the users who tweeted about Joe Biden's presidency (Topic 8), we find that these users had a definitive rightward tilt (+0.5491). This tilt is also observed for those who tweeted about the Biden presidency in a toxic manner (+0.300 rightward tilt). 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). Finally, examining the set of users who tweeted about the raid of Donald Trump's home for classified documents (Topic 10), we again see a rightward bias (+0.0656), with an even stronger right lean among users who addressed it in a toxic manner (+0.0797).

Besides these politically salient issues, we observe several topics where politically charged users simply derided each other (Topic 1), called each other racist (Topic 5), or insinuated that the other political side supports pedophilia (Topic 6). Furthermore, as seen in Table 3, a particular epithet utilized by right-leaning users against US President Biden was "PedoHilter" (Topic 6). As reported elsewhere [77], this hashtag and phrase trended on Twitter after US President Biden's "Speech on Saving Democracy" on September 1, 2022.

5.3 The Most Partisan Toxic Topics of 2022

Having explored the set of topics with the most toxic tweets, we now examine the set of most right-leaning and left-leaning topics within our dataset. As seen in Tables 4 and 5, in the several

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	trump, sides, bullsh*t, country, democrat	4,554	333 (7.31%)	0.151	Trump Is A Criminal Trump Is A Traitor Trump Is Going To Jail	-1.200	-0.320	1.253
2	badd, gawd, fuqqing, motherfucker, f'ing	22,345	450 (2.01%)	0.116	@REDACTED Hes nuts and the Obama administration was right to boot him	-1.119	-1.10	0.773
3	twiddle, lol, gop, baked, dumb	5,417	184 (3.40%)	0.087	@REDACTED @REDACTED "Flooding the zone with shit"	-1.090	-0.883	0.935
4	she, acosta, asshat, tucker, mtg	31,042	3450 (11.11%)	0.241	He's a sociopath just like Abbott, Cruz and the rest of them.	-0.994	-0.836	0.838
5	sentence, bannon, jail, capitol, lock	21,287	1361 (6.39%)	0.138	@REDACTED If it looks like shit, and smells like shit, it's probably Bannon. #AccountabilityNow	-0.962	-0.787	0.935

Table 5. Top toxic topics with at least 100 toxic tweets by left-leaning tilt.

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	youngkin, virginian, governor, glenn, gop	16,976	1024 (6.03%)	0.100	Dear Virginia, You went from blue to stupid red and racist in one day. Great job!	-0.376	-0.130	1.515
2	romance, caribbean, ransom, extortion, virgin	5,084	119 (2.34%)	0.055	In which James Max exposes the ridiculous, laughable ideas behind gender identity ideology.	-0.596	-0.0966	1.324
3	thehill, bull, utterly, ridiculous, headline	41,509	3251 (7.83%)	0.085	@thehill Hahahahahahaha, you lying scum	0.221	0.182	1.267
4	mcmullin, evan, mike, utah, mcmuffin	2,995	215 (7.18%)	0.118	Evan Mcmullan is just a straight up asshole	0.0857	0.296	1.266
5	joy, msnbc, joyless, miserable, television	2,833	145 (5.12%)	0.176	@Joy has turned into a ridiculously naive half-witted ignoramus.	-0.0716	0.697	1.260

Table 6. Top toxic topics with at least 100 toxic tweets by variation in user political ideology.

cases, many of the most right-leaning and left-leaning topics simply disparage the other political ideology (Topic 1 and 3 in Table 4; Topic 1, 3, and 4 in Table 5). For example, the most right-leaning topic calls members of the US Democratic party "DemoKKKrats", while the most left-leaning topic calls former Republican US President Trump a criminal and calls for his arrest. In addition, we further observe that some of the most right and left-oriented topics are replies/mentions to specific users (Topic 2 in Table 4; Topic 1, 2, 4 in Table 5). Examining each of these in turn, we find that, in many cases, these topics are either attacks on particular political officials or replies to tweets of other hyperpartisan users. Similarly, examining the set of topics with the widest variation in user's political orientation (Table 6), we again see that nearly all of them are tweets targeted at political leaders or TV personalities (*e.g.*, Glenn Youngkin, Evan McMullin, Joy Reid) with some users defending them and others attacking them vehemently.

Right-Leaning Topics. Beyond the most partisan right-leaning topics deriding the left as "DemoKKKrats" (Topic 1) and "fuckers" (Topic 3), among the top right-leaning topics, we observe a topic aimed at promoting the *Gutfeld!* FoxNews talk show mocking the CBS talk show *The Late Show with Stephen Colbert*. A conservative talk show *Gutfeld!*, in August 2022, eclipsed the left-leaning *The Late Show with Stephen Colbert*, as the most popular comedy-focused late-night television program [45].

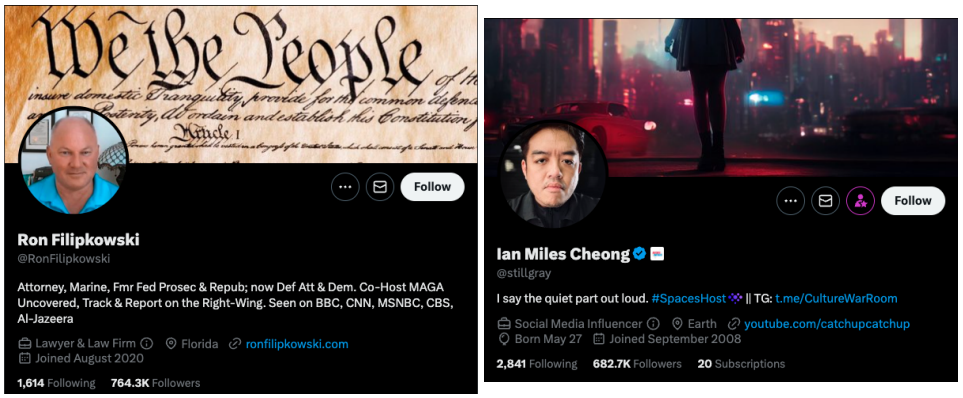


Fig. 11. Accounts like @ronfilipkowski and @stillgray had users of similar political orientations reply in a toxic manner to the news and opinions they tweeted.

Account	# Campaigns	Avg Partisanship of Campaigns
@youtube	14	0.076
@elonmusk	9	0.065
@laurenboebert	8	-0.242
@thehill	7	0.069
@ronfilipkowski	7	-0.823
@atrupar	7	-0.368
@REDACTED	7	-0.247
@stillgray	5	0.399
@gbnews	4	0.006
@hawleymo	4	-0.027

Table 7. Number and average partisanship of toxic reply/mention campaigns encountered by various Twitter accounts.

Besides the instances of topics targeting particular Twitter users, we lastly observe a topic of right-leaning users maligning supposed RINOs (Topic 5). Republicans In Name Only or RINOs, largely indicating moderate or establishment Republican officials, have become a target of ire throughout hyper-conservative circles [88] as further seen in our dataset.

Left-Leaning Topics. Many of the most liberal oriented-topics target the Republican Party (Topic 3), current or former US Republican public officials or conservative media figures including former US President Trump (Topic 1), former Trump advisor Steve Bannon (Topic 4), conservative commentator Tucker Carlson (Topic 5), and Texas Governor Greg Abbott (Topic 4).[11] Most notably, Topic 5 concerns focuses on the news story when Steve Bannon was sentenced to four months in prison after he refused to turn over documents subpoenaed by the US House in their investigation of the attack on the US capitol on January 6, 2021 [44].

"Toxic Topics Campaigns." We note that by examining the rest of our clusters and as also observed in Table 6, we observe several other separate instances when various users encountered "toxic topics campaigns" of toxic replies/mentions (*i.e.*, where the majority of tweets were "@"s at a particular account). For example, while we displayed one just toxic campaign targetting @GlennYoungkin, we identified two other such campaigns in our dataset. Altogether we identified 1,923 campaigns

[11]Note that we do not discuss Topic 2 the replies are to a hyper-left-leaning redacted account

against 1,852 users. 573 of these campaigns have a right-leaning orientation (*i.e.*, average political ideology of campaign participant > 0) while 1,350 have a liberal orientation. Calculating the political orientation of these "attacked" accounts, across these campaigns, 15.5% were in cases of right-leaning accounts campaigning against liberal accounts; 17.4% were cases of liberal accounts campaigning against right-leaning accounts; 29.2% were right-leaning against right-leaning; 37.8% were left-leaning against left-leaning. Compared to all mentions where only 30.8% are between users of different political orientations, we thus again observe evidence of affective polarization in these "toxic topics campaigns." In Table 7, we present the number of "toxic topics campaigns" against particular users. 29.2% (541 accounts) of the "attacked" accounts were verified (compared to only 9.3% [5,157 accounts] of the accounts out dataset of 55,415 Twitter users), suggesting that more public figures are more likely to incur these campaigns.

We note that, while in some cases these are targeted campaigns meant to attack particular users, in several cases these toxic campaigns are other Twitter accounts toxicly responding in agreement to the opinions or news put forward by the account. While the campaigns targeting @laurenboebert, a conservative congressperson from Colorado, are mostly by heavily left-leaning users for example, this occurs in reverse for the user @stillgray, a hyperpartisan conservative commentator, and @ronfilipkowski, a hyperpartisan liberal commentator. Other campaigns, for instance, were aimed at @YouTube to protest particular videos being taken down. We leave it to future work to fully explore and differentiate between these types of "toxic topics campaigns."

5.4 Topic Dependent Changes in Political Polarization and Toxicity

Having detailed many of the most toxic and partisan topics within our dataset, we now explore how the toxicity of conversations changes 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 $\rho = -0.029$, indicating little to no relationship. Furthermore, across our dataset, we find that regardless of whether the topic moderates or moves to the extremes, in both cases, toxicity generally increases (55.4% of the time for topics that moderated in political ideology and 50.0% of the time for topics that moved to the political extreme). Furthermore, we find that across our dataset between January 2022 and December 2022, in 34.7% of topics, as topics became more right-leaning, they also became more toxic; in 27.5% cases, they became less toxic as they became more right-leaning. Conversely, in 20.5% of our topics, they became more toxic as they became more left-leaning, and in 17.2% of topics they became less toxic as they became more left-leaning. However, examining each cluster, we *do* find that on a cluster-by-cluster basis as the political composition of users involved in that topic changes there are corresponding changes in toxicity.

Plotting toxicity and political orientation over time for the topics with the largest increases in toxicity between January 2022 and December 2022, we observe that while for four topics considered, (Figures 12a, 12b, 12d, and 12e) as the topic became more right-leaning, toxicity similarly increased, for one of the topics (Figure 12c), 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 opposite direction, is correlated with an increase in toxicity. For instance, in the tweets surrounding current US President Joe Biden's son, Hunter Biden, we observe that as users discussing Hunter Biden became more right-leaning, the more toxic the tweets became. Hunter Biden was investigated by the US Federal Bureau of Investigation (FBI) and subsequently was given two misdemeanor charges related to his 2017 and 2018 taxes and for a separate charge the

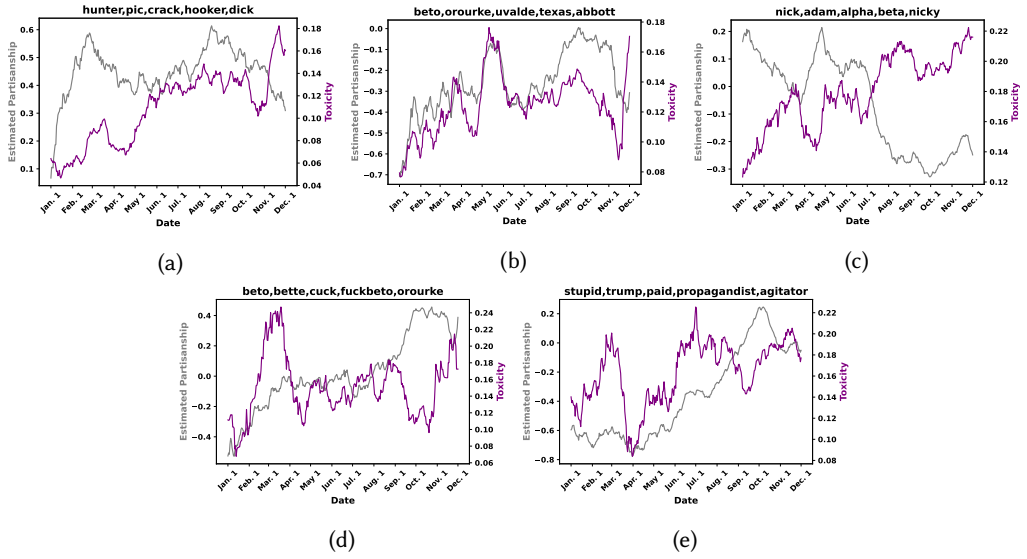


Fig. 12. Topics with the largest increase in toxicity in 2022.

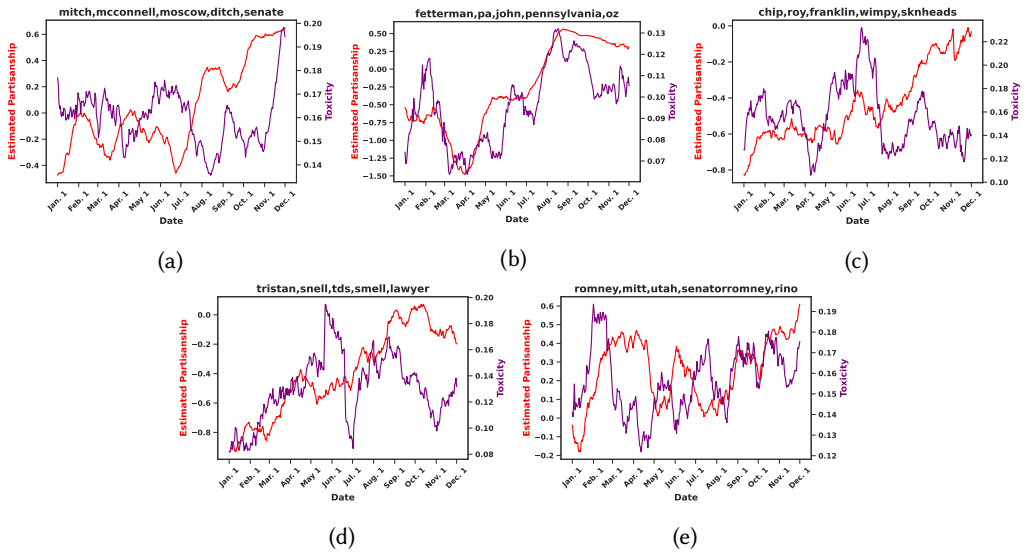


Fig. 13. Topics with the largest swing to right-leaning partisanship throughout 2022.

unlawful possession of a handgun [49]. Given the highly political nature of these charges and their use within right-leaning media to level attacks against the Democratic US President Joe Biden, the toxic and right-leaning nature of this topic is unsurprising. Similarly, for two topics (Figures 12b, 12d) that centered around the Democratic Texas politician and former Texas gubernatorial O'Rourke, as the topic became more right-leaning, the corresponding topic became more toxic. Finally, examining the conversation that became more toxic as it became more left-leaning, we observe that it centers around the conservative commentator Nick Adams, founder of the *Foundation for Liberty and American Greatness* [89]. We thus observe among these top topics that depending on the political

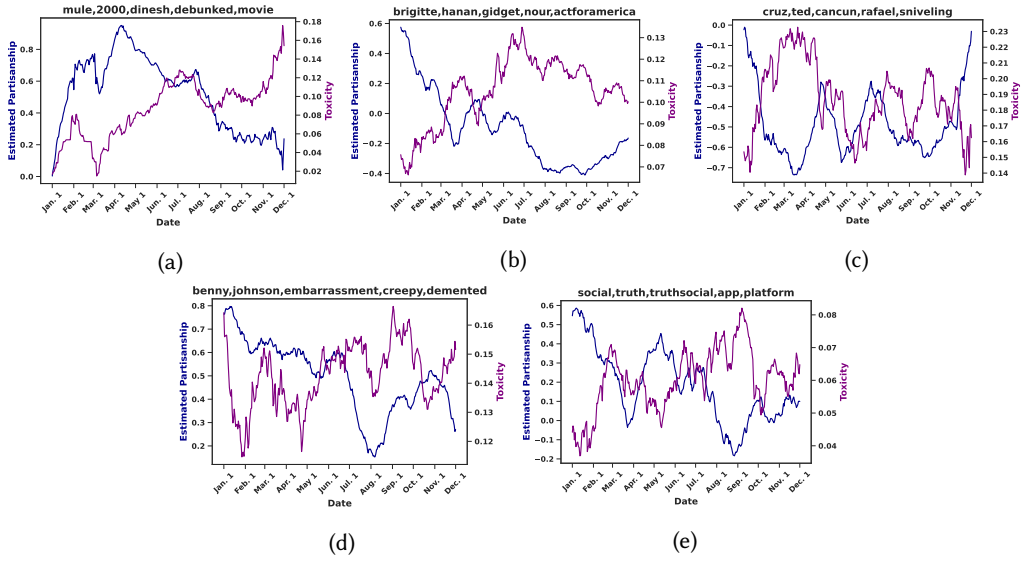


Fig. 14. Topics with the largest swing to left-leaning partisanship throughout 2022.

nature of the given topic, a corresponding swing in the political composition of the users in the opposite direction, may be correlated with an increase in toxicity.

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 14 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 Chip Roy (the Republican representative for Texas' 21st congressional district) became more right-leaning, the toxicity of that topic decreased dramatically (Figure 13c). Similarly, as liberal-leaning users began to join the conversation surrounding Benny Johnson (a conservative commentator for the right-wing Newsmax media outlet) and Dinesh D'Souza's movie *2000 Mules* about supposed voter fraud in the 2020 US Presidential election [13], the conversation also became more toxic (Figures 14d, 14a). Conversely, we find that as right-leaning users joined the conversation about US Senate Republican Minority Leader Mitch McConnell being beholden to the Russian government [42] and Utah Senator Republican Mitt Romney being a RINO [79], these topics increase in toxicity (Figures 13a, 13b 13e). We note that the attacks against Senators Mitch McConnell and Mitt Romney were largely for not being conservative enough. 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. We thus conclude that political ideology itself does not necessarily predict a higher degree of toxicity within conversations (as with users [see Section 4.3]), but is largely topic-dependent, with even the topic being a right-leaning or left-leaning entity/individual not decisively giving whether left or right leaning shift in users will corresponding to an increase in toxicity.

5.5 Topic User Composition and the Toxicity of Topics

Having examined the composition and changing dynamics 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.

Adjusted R-squared: 0.373	Coefficient	Std. Error	Adj. Sum Sq.
Intercept	-0.108^{***}	0.012	0.983
Log # Users	-0.005	0.003	0.029
Percentage Verified	0.040	0.020	0.015
Avg User Toxicity in Cluster	$\mathbf{2.897}^{***}$	0.065	**16.95**
Abs Cluster Political Ideology	-0.093^{***}	0.015	0.352
Std Cluster Political Ideology	0.0415^{***}	0.010	0.072
Left/Right	0.0054	0.003	0.026

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 8. Linear fit on factors in the toxicity in individual topic clusters.

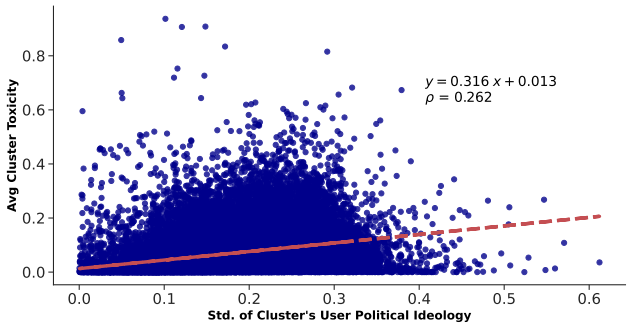


Fig. 15. As previously also found in our analysis of user characteristics (Section 4.3), we find that as users engage in a wider window of topics of particular political ideologies, the more toxic their tweeted content

We note, and as seen throughout this section, topics on Twitter vary widely with individual topics often varying widely in political composition over time. Topics and the users addressing them change dramatically throughout the year. 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 0.168 standard deviations (based on the latent space that we previously determined [Section 3.1]). In 62.0% of cases, topics became more right-leaning, and in 38.0% topics became more left-leaning; similarly, within this same period, 58.1% became more toxic while 41.8% became less toxic. As a result, to quantify the effect that the composition of users has on the toxicity of a given topic, for this section, we limit our analysis to the tweets posted within a single month, December 2022 (*i.e.*, the political composition of a given topic's user and the toxicity of that cluster are more stable).

To determine the role of various topic-level features in the overall toxicity of that cluster, we fit a linear regression on the average toxicity score within each of our clusters against

(1) The number of users who tweeted about that topic,
(2) Te average user toxicity in the cluster
(3) The percentage of users involved in that topic that is Twitter verified
(4) The absolute value of the political ideology in that cluster
(5) the standard deviation of political ideologies of users within that topic cluster.
(6) Whether that cluster has a leftward or rightward political lean.

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). We note that we again reproduce our results with the Perspective Toxicity API in Appendix F obtaining similar results. As seen in Table 8, 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 covariate accounts for 34.31% of the variation in 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 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.

We further find, however, that as conversations become more politically diverse, with more types of users becoming involved, that toxicity increases. This again reinforces the presence of affective polarization on Twitter. Looking in the reverse direction, we further confirm, as seen in Figure 15 that as users are involved in a higher variance of topics of different political topics their average toxicity increases as well. We thus find from this analysis further confirmation, on a topic level,l that increased user toxicity and the diversity of views present in a given conversation contribute to toxicity within particular topics. We now consider some of the implications of these results.

6 DISCUSSION

In this work, we considered factors that contribute to 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 in the political center have a higher likelihood of tweeting toxic messages. We similarly find that users who interact with other users who more regularly post toxic content are more likely to post toxic content themselves. Further, as users interact with or mention other users from a wider range of political ideologies, they are more likely to post toxic content.

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 is likely to be. Here we discuss some of the limitations and implications of our results:

6.1 Limitations

We acknowledge several limitations of this work. Given our use of linear regressions to estimate the effect of partisanship and political diversity, our findings are largely correlational. While they do buttress and support a large literature of similar results [6, 8, 19, 52], we acknowledge that our results are not causal. We further note that due to new restrictions placed on the collection of Tweets [75], we can not continue and measure the toxicity of users and political topics, going forward.

We further note that 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.

Finally, 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 [53, 68, 73], we do base our DeBERTa-based model toxicity detection on this definition, and thus we acknowledge that it may not take into account all perspectives on what constitutes toxic online content.

6.2 Tribal Tendency, Affective Polarization, and Online Toxicity

As found by others political heated conversations often elicit toxicity as people of differing views debate and discuss their differences [70]. On Twitter, we find, that this discourse is related to increased toxicity. Political diversity, at least in the short form of tweets, is correlated with affective polarization and toxic content. While users naturally often congregate and more heavily engage

with users like themselves (assortativity coefficient of 0.258), when they do engage with other users of differing political views, we observe that this tends to create conflict. While this feature of online conversation is not the dominant factor in engendering toxic content, with other factors like a user's previous behavior [56], 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, illustrating the robustness of this finding. We further note that this finding reinforces De Francisci Morales *et al.* [21] finding in Nature that interactions among users on Reddit with different political orientations increased negative conversational outcomes.

6.3 Hyperpartisan Users and Topics

Throughout this work, we found 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 ($\rho = 0.594$) and as a result, do not necessarily have higher toxicity levels. 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 (*e.g.*, topics and users nearer to the political center). This result, somewhat unexpected, indicates that political echo chambers, where only left-leaning or right-leaning interact amongst themselves, are less conflict-oriented on Twitter.

6.4 Intra-Topic Political Ideology over Time

In Section 5.4, we observed that the political orientation of users that discuss any particular topic often changes over time. These changes, often coinciding with changes in toxicity, also illustrate the views expressed on Twitter about particular topics often change throughout the year as different users enter or leave different conversations. We argue that future analysis of topics and their spread on Twitter *must* take into account user-level characteristics such as political ideology given that these values often reveal the nature of how users are addressing individual topics. For example, as seen in Section 5.4, 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 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 [50, 54, 74]. While not a new finding [52], this illustrates 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 [68], where toxicity is 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 Future Work

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

Identifying Hate and Toxic Topic Campaigns As seen in Section 5.3, utilizing our approach, we managed to identify various instances where accounts encountered toxic tweets aimed at them and centered around a particular topic. For instance, we identified 16,976 tweets (1,024 toxic) aimed at Virginia Governor Glenn Youngkin; altogether we identified 1,923 similar "campaigns" aimed at 1,852 different accounts. We note that our DeBERTa-based model, which we open-source for study and use, can further enable others to continue this work without having to rely on making online and black-box queries to the Perspective API. In future work, we plan to better identify when users specifically attack particular accounts by training a model to predict the "ATTACK ON AUTHOR" task provided by Google Jigsaw [1] along with other metadata embedded within Twitter conversations. While not the focus on this particular work, we note that by being able to automatically identify potential "toxic" campaigns against particular users, our work could be utilized to protect journalists, public officials, and vulnerable populations.

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, *etc...*) to identify the role of partisanship and political polarization.

7 CONCLUSION

In this work, we analyze the role that a variety of different factors have in contributing to potentially why users post toxic content and why conversations themselves are toxic. Analyzing 187 million different tweets from 55,415 users from across the political spectrum, we find a user or topic being heavily partisan does not necessarily imply increased toxicity; rather as users engage with and as conversations involve a wider range of political orientations and with other toxic users and toxic content that online toxicity increases.

REFERENCES

[1] 2022. Google Jigsaw. Perspective API. https://www.perspectiveapi.com/#/home.

[2] Ana Aleksandric, Sayak Saha Roy, and Shirin Nilizadeh. 2022. Twitter Users' Behavioral Response to Toxic Replies. *arXiv preprint arXiv:2210.13420* (2022).

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

[4] 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.

[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] 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.

[7] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.

[8] 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.

[9] 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.

[10] Porismita Borah. 2013. Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication* 30, 3 (2013), 456–473.

[11] 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.

[12] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009).

[13] Kate Brumback. 2023. Georgia sues True the Vote group over refusal to produce evidence of '2000 Mules' claims. *Associated Press* (7 2023). https://www.fox5atlanta.com/news/georgia-sues-true-the-vote-produce-evidence-mules-documentary-claims

[14] 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.

[15] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603* (2020).

[16] 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.

[17] 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.

[18] Dana Cuomo and Natalie Dolci. 2019. Gender-Based Violence and Technology-Enabled Coercive Control in Seattle: Challenges & Opportunities.

[19] Chrysi Dagoula. 2019. Mapping political discussions on Twitter: Where the elites remain elites. *Media and Communication* 7, 1 (2019), 225–234.

[20] 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/

[21] 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.

[22] 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*.

[23] 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.

[24] Maeve Duggan. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[25] 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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

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

[33] 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.

[34] 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.

[35] Hans WA Hanley and Zakir Durumeric. 2023. Partial Mobilization: Tracking Multilingual Information Flows Amongst Russian Media Outlets and Telegram. *arXiv preprint arXiv:2301.10856* (2023).

[36] 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).

[37] 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).

[38] 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*.

[39] 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.

[40] 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.

[41] 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.

[42] 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

[43] 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.

[44] Carrie Johnson. 2022. Steve Bannon sentenced to 4 months in prison : NPR. https://www.npr.org/2022/10/21/1130327514/steve-bannon-sentencing-jan-6-committee

[45] Ted Johnson. 2022. Fox News Tops August Ratings And 'Gutfeld!' Has Highest-Rated Month – Deadline. https://deadline.com/2022/08/fox-news-august-ratings-gutfeld-1235103363/

[46] Andreas Jungherr. 2014. Twitter in politics: a comprehensive literature review. *Available at SSRN 2402443* (2014).

[47] Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph. D. Dissertation.

[48] 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.

[49] Thomas Kika. 2023. Hunter Biden Scandal 'Extremely Serious,' Former FBI Agent Says. https://www.newsweek.com/hunter-biden-scandal-extremely-serious-former-fbi-agent-says-1813177

[50] 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.

[51] Brian Kulis and Michael I Jordan. 2011. Revisiting k-means: New algorithms via Bayesian nonparametrics. *arXiv:1111.0352* (2011).

[52] 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.

[53] 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.

[54] 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).

[55] 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).

[56] 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.

[57] 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.

[58] Alice Marwick and Danah Boyd. 2011. To see and be seen: Celebrity practice on Twitter. *Convergence* 17, 2 (2011), 139–158.

[59] 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.

[60] 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

[61] 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.

[62] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.

[63] Keith T Poole and Howard Rosenthal. 2007. On party polarization in Congress. *Daedalus* 136, 3 (2007), 104–107.

[64] Walter Quattrociocchi, Rosaria Conte, and Elena Lodi. 2011. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems* 14, 04 (2011), 567–586.

[65] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).

[66] 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.

[67] 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.

[68] 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.

[69] 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).

[70] 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.

[71] 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.

[72] 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.

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

[74] 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.

[75] 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/

[76] 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).

[77] Staff. 2022. President Biden's Speech About Saving Democracy Angers Right-Wing Extremists, Politicians | ADL. https://www.adl.org/resources/blog/president-bidens-speech-about-saving-democracy-angers-right-wing-extremists

[78] 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*.

[79] Paul Steinhauser. 2020. Trump calls Romney a 'RINO' after GOP senator targets president's 'undemocratic action'. *Fox News* (11 2020). https://www.foxnews.com/politics/trump-romney-rino-president-undemocratic-action

[80] Cass R Sunstein. 2018. Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.* 27 (2018), 83.

[81] Edson C Tandoc Jr and Erika Johnson. 2016. Most students get breaking news first from Twitter. *Newspaper research journal* 37, 2 (2016), 153–166.

[82] 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.

[83] 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.

[84] 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).

[85] 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.

[86] Peter D Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on machine learning*.

[87] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[88] Dylan Wells and Colby Itkowitz. 2023. Policy demands, personal animus and more: Meet the McCarthy resistance - The Washington Post. https://www.washingtonpost.com/politics/2023/01/05/mccarthy-critics-house-republicans-speaker/

[89] Luke Winkie. 2023. MAGA Twitter: Pro-Trump accounts have reached a strange new low. https://slate.com/human-interest/2023/02/donald-trump-twitter-nick-adams-maga.html

[90] 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.

[91] 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.

[92] 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.

A CORRESPONDENCE ANALYSIS FOR APPROXIMATING POLITICAL IDEOLOGY

After identifying our set of 882 politically discriminating 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 run the 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 followed by the conservative and liberal accounts in the first stage of our analysis most often. As in Barbera et al [6], 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 2 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 further can estimate the political ideology of any account that follows at least one of 1607 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...*). Altogether we project an additional 49,308 users.

B POINTWISE MUTUAL INFORMATION

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

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

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 [86].

C DP-MEANS

DP-Means [51] is a non-parametric 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 a chosen

parameter λ away from the closest cluster, a new cluster is formed. Dinari *et al.* [22] 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 instead 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 datapoints create new clusters) [22].

D LINEAR FIT OF OF USER-LEVEL FEATURES AND PERSPECTIVE TOXICITY

Adjusted R-squared: 0.307	Coefficient	Std. Error	Adj. Sum Sq.
Intercept	0.0735	0.003	3.444
Verified Status	-0.0103	0.001	0.253
Years Active on Twitter	-0.007	-0.00008	0.438
Log # Followers	-0.0132	-0.001	2.056
Log # Followed	-0.0132	-0.0014	1.285
Log # Tweets in 2022	0.0111	0.0145	1.960
Toxicity of Users in Mentions	**0.8527**	0.012	**27.901**
Abs Political Ideology	-0.0075	0.001	0.676
Left/Right	0.0031	0.001	0.094
Std. of Mentioned Users Political Ideology	0.0287	0.001	2.210
Avg Abs(User Ideology - Mentioned Users Political Ideology)	0.0310	0.001	3.222
Abs Avg of Mentioned Users Political Ideology	-0.0247	0.001	1.721

Table 9. Linear fit of several user-level factors on average toxicity of Twitter users. All coefficients had a p-value ≈ 0 using t-tests. We perform Analysis of Variance (ANOVA) tests to further determine significance and estimate the variation that can be attributed to various factors (Running ANOVA we see that all coefficients are indeed significant and contribute to explaining toxicity on Twitter).

E MOST TOXIC TOPICS BY PERCENTAGE

Topic	Keywords	# Tweets	# Toxic Tweets	Avg. Toxicity	Example Tweet	Avg. Partisan.	Avg. Partisan of Toxic Users	Partisan Std.
1	fuck, shit…, shit, shittttt, extremely	165	117 (70.91%)	0.896	This was my shit!!!!!!	-0.0565	-0.103	0.220
2	stupid, rhetorical, special, really, donald	884	486 (54.98%)	0.892	How stupid do you have to be?	0.241	0.175	1.006
3	herself, yourselves, himself, go, fuck	518	223 (43.05%)	0.888	The guy who should go fuck himself has thoughts.	0.1367	0.011	0.955
4	idiot, useful, calling, idiotic, blithering	19604	10589 (54.01%)	0.886	I believe he qualifies as an Idiot!!	0.187	0.006	0.9417
5	full, piece, shithead, shit, pile	621	549 (88.41%)	0.881	Pieces of SHIT are “birds of a feather”.	-0.0466	-0.192	0.912

Table 10. Top toxic topics—by average toxic value—in our dataset.

F LINEAR FIT OF TOPIC-LEVEL FEATURES AGAINST PERSPECTIVE TOXICITY

Adjusted R-squared: 0.188	Coefficient	Std. Error	Adj. Sum Sq.
Intercept	0.1001^{***}	0.012	0.850
Log # Users	-0.0117^{***}	0.003	0.165
Percentage Verified	-0.255^{***}	0.038	0.559
Avg User Toxicity in Cluster	5.42^{***}	0.219	**7.633**
Abs Cluster Political Ideology	-0.032	0.019	0.037
Std Cluster Political Ideology	0.134^{***}	0.019	0.627
Left/Right	0.0144^{***}	0.004	0.181

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 11. Linear fit on factors in the toxicity in individual topic clusters.