Sub-Standards and Mal-Practices: Misinformation's Role in Insular, Polarized, and Toxic Interactions

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

How do users and communities respond to news from unreliable sources? How does news from these sources change online conversations? In this work, we examine the role of misinformation in sparking political incivility and toxicity on the social media platform Reddit. Utilizing the Google Jigsaw Perspective API to identify toxicity, hate speech, and other forms of incivility, we find that Reddit comments posted in response to articles on websites known to spread misinformation are 71.4% more likely to be toxic than comments responding to authentic news articles. Identifying specific instances of incivility and utilizing an exponential random graph model, we then show that when reacting to a misinformation story, Reddit users are more likely to be toxic to users of different political beliefs. Finally, utilizing a zero-inflated negative binomial regression, we identify that as the toxicity of subreddits increases, users are more likely to comment on misinformation-related submissions.

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: Misinformation, Toxicity, Political Polarization, Reddit, Online Communities

1 INTRODUCTION

Over the last decade, misinformation, incivility, and political polarization have corroded trust in democratic institutions [15, 21, 46, 47, 50]. While separate and distinct phenomena, misinformation, toxic language, and political polarization often combine with one another, stoking division and negatively affecting social media platforms [17, 25, 29, 31, 49, 82, 106, 112, 116]. While several works have attempted to understand the impact of these individual factors, in this work, we explore their relationships with one another, asking the following three research questions:

(1) *Do posts of articles from misinformation sites on social media lead to increased toxicity in comments compared to articles from authentic news submissions?*
(2) *What role do differences in political ideology play in increased toxic interactions (i.e., affective polarization) between users in the presence of misinformation?*
(3) *Do the toxicity and political ideological norms of an online community change the amount of interaction that users have with misinformation and unreliable news?*

To answer these questions, we measure the levels of toxicity, political ideology, and misinformation within communities on the social media platform Reddit for an 18 month period (January 2020 to June 2021). Specifically, we determine the number of toxic comments within each community and by individual users on Reddit using the Google Jigsaw API [2], a commonly deployed classifier for identifying toxic language (*e.g.*, insults, sexual harassment, and threats of violence [110]). Then, utilizing a hyperlink-based approach as outlined by Saveski *et al.* [100], we approximate the political orientations of a subset of subreddits (Reddit communities) and users along the US left-right political spectrum. Finally, we utilize previously curated lists of misinformation websites to determine the levels at which these communities and users post links to websites known to spread misinformation online. From these calculations, we tackle our three research questions about the relationships between toxicity, political ideology, and misinformation:

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.

RQ1: Toxicity and political ideology in misinformation posts. Utilizing our list of misinformation outlets, a separate list of authentic news websites, and our measured toxicity and political norms of users and subreddits, we first determine whether there are distinct levels of user political partisanship and toxicity in the comments posted in response to articles from misinformation and authentic news websites. We find that comments on articles from misinformation websites are toxic at a rate 71.4% higher than comments in response to authentic news (1.80% of comments on misinformation posts are toxic versus 1.05% of comments to authentic news). We further observe that users that comment under submissions that reference articles from misinformation sources tend to be slightly more right-leaning than those that do not. Finally, examining misinformation's correlation with overall toxicity, we find that as levels of misinformation increase in a subreddit, the average toxicity of comments is also higher ($\rho = 0.352$). We confirm these results by fitting a linear regression model against the toxicity of individual submissions that accounts for each of these features.

RQ2: Misinformation's correlation with inter-political strife. Having identified that users who comment on misinformation posts are more likely to make toxic comments than those who respond to authentic news posts, we examine the role of political ideology in these toxic interactions. We observe that subreddits with higher amounts of *misinformation-oriented* posts are more likely to have more intra-party and insular interactions relative to subreddits with more *authentic news-oriented* posts. Utilizing an exponential random graph model, we further find that compared to other Reddit users, users who comment on misinformation posts are more likely to respond to users of different political views in a toxic manner.

RQ3: Toxic subreddits and engagement with misinformation. Lastly, having documented the role of misinformation in inciting toxicity, especially among users of different political orientations, we determine how user toxicity and political ideology predict *community engagement* with misinformation and authentic news. Fitting a zero-inflated negative binomial model to our data, we find that as subreddits become more toxic and more politically ideologically extreme, Reddit users are more likely to comment on misinformation news article submissions. This contrasts with authentic news submissions, where more toxic communities are less likely to engage with mainstream articles.

Altogether, in this work, we document misinformation's role in politically insular and toxic communities and in predicting toxic interactions between users. Our work, one of the first to examine the relationship between misinformation, toxicity, and political ideology, illustrates the need to fully understand the complex interactions between these phenomena so that platforms can better understand and address toxicity online.

2 BACKGROUND & RELATED WORK

In this section, we detail several key definitions, provide background on Reddit, and present an overview of prior works that analyze the effects of misinformation, toxicity, and political polarization on social media.

2.1 Terminology

The role of social media in promoting misinformation-heavy, toxic, and highly politically polarized ecosystems has been intensely studied [24, 52, 58, 110]. Utilizing these studies, we first provide several key definitions that we use to ground our own work.

Misinformation and Authentic News. As in previous studies, we define *misinformation* as information that is false or inaccurate regardless of the intention of the author [8, 53, 58, 65, 70, 79, 119]. Similarly, we define *misinformation websites* or "unreliable sources" as news websites that

regularly publish false information or misinformation about current events and that do not engage in journalistic norms such as attributing authors and correcting errors [4, 8, 22, 58, 62, 86, 104, 125]. Conversely, we define *authentic news websites* as news websites that generally adhere to journalistic norms including attributing authors and correcting errors; altogether publishing mostly true information [58, 62, 125].

Online Toxicity and Incivility. Given our use of the Google Jigsaw Perspective API [2], we use their definition of online toxicity/incivility throughout our work. Namely, toxicity is "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion.*"

Political Ideology/Partisan Bias. We define political ideology/partisan bias as users' and communities' place on the US left/right political spectrum [97]. We note the limitation of this definition given the variety of political views within the US. However, in line with previous work [59, 99, 100], we utilize this definition, which largely fits much of US-centered political discussions, in order to understand how right-leaning and left-leaning users and communities interact with one another and misinformation.

Affective Political Polarization: Affective political 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 has become a defining aspect of US politics [32].

2.2 Reddit

Reddit is an online social media platform composed of millions of subcommunities known as subreddits [3, 19]. Subreddits are each dedicated to specific topics, ranging from politics (r/politics) and science (r/science) to Pokemon (r/pokemon). Depending on their community guidelines and rules, users can submit news articles, opinions, images, and memes as *submissions*. Underneath these submissions, other users can leave comments or reply to comments from other users. Anyone can create a subreddit and these subreddits are moderated both by Reddit content policies, subreddit-specific rules, and implicit community norms [19, 37, 69]. Weld *et al.* [120] find that subreddit norms can vary widely. These norms encompass political behaviors, tolerance to misinformation, and toxic behavior [19, 69, 94, 120].

2.3 Political Ideology and Polarization

People, both in real life and on the Internet, tend to associate with like-minded people [10, 11, 54, 56, 66, 72, 90]. Wojcieszak *et al.* [122] find that while the majority of political discussions online are between participants that share the same viewpoint, many users *do* enjoy conversations with people with different viewpoints [108]. Social media can thus have the benefit of exposing individuals to multiple views by allowing users to interact with different types of people [11, 30, 90]. Despite this potential, past works have found that many social media platforms are one of the main reasons for high degrees of political polarization across the globe [17, 18, 60, 72]. Cass Sunstein, Garett *et al.*, and Quattrociocchi *et al.* all argue that the "individualized" experience offered by social media platforms comes with the risk of creating "information cocoons" and "echo chambers" that accelerate polarization [45, 91, 109]. Conover *et al.* [27], for example, find that different structures of conversations on Twitter interactions are often heavily influenced by Twitter's own structure fostering increased levels of politically polarized conversations. Bessi *et al.* [14], examining the behaviors of 12 million users, find that partisan echo chambers are driven by the algorithms of both Facebook and YouTube. Torres *et al.* [111] find the specific Twitter behavior of "follow trains" induce highly politically polarized behavior on the platform.

In a similar vein, prior work has found that the increased political polarization engendered by social media causes several negative downstream effects including the increased sharing of

misinformation and toxic online behaviors. Imhoff *et al.* [68], for example, find that political polarization is associated with beliefs in conspiracy theories. Ebling *et al.* [34] similarly find that political partisanship levels on social media are associated with medical misinformation about COVID-19. Other studies have further interrogated the adverse effects that social media has had on the democratic process due to the increased political polarization associated with social media [51, 88, 112, 113].

2.4 Misinformation

In addition to driving political polarization, online activity has been found to be one of the main drivers of misinformation. As researched and reported extensively, misinformation has increasingly become a major and distinctive aspect of the conversations on social media [8, 43, 48]. Even after controlling for cascade size, Juul and Ugander find that false information spreads deeper and wider on Twitter than true information [71]. Furthermore, misinformation often convinces those that are exposed to it. A large percentage of US adults were exposed to misinformation stories by social media during the 2016 election [8] and many believed these false stories were true [7, 53]. As COVID-19 spread throughout the world, online misinformation and conspiracy theories became a major hurdle to curbing its spread [98, 105].

To prevent the spread of misinformation, recent research has heavily focused on tracking and stemming its flow [58, 112]. For example, Mahl *et al.* [81], track the spread of 10 different conspiracy theories on Twitter, identifying one of the largest conspiracy theorist networks. Ahmed *et al.* [5] use a similar approach to track the spread of COVID-19 and 5G conspiracy theories. They find well-known misinformation websites were some of the largest sources helping to spread these conspiracy theories on Twitter's platform. Gruzd [52] found that a single Tweet about how COVID-19 was a hoax, spanned an entire conspiracy theory, eventually prompting large groups of people to film their local hospitals to prove that COVID-19 was not real. In addition to network-based approaches, several others have used advancements in natural language processing to identify and track misinformation. Hanley *et al.* [57], for example, utilize semantic search to identify and track Russian state-media narratives on Reddit. Fong *et al.* [39] utilized linguistic and social features to understand the psychology of Twitter users that engaged with known conspiracy theorists on the Twitter platform. Finally, several works have performed in-depth case studies on the spread of specific misinformation narratives. In their papers, Wilson and Starbird *et al.* look at the Syrian White Helmets on Twitter, and Bär *et al.* look at the spread of QAnon on Parler [16, 121].

2.5 Toxicity

41% of Americans and 40% of those globally have reported experiencing bullying or harassment online [33, 110]. Online toxicity takes many forms including threats, sexual harassment, doxing, coordinated bullying, and political incivility [41, 42, 80, 110]. Toxic comments, in particular, are one of the most common forms of hate and harassment online [110]. Similar to our definition (Section 2.1), Vargo *et al.* [115] describe toxic comments as those that utilize "extremely vulgar, abusive, or hurtful language". Muddiman *et al.* define online political toxicity [84] as comments that violate "politeness norms, such as name-calling and swearing, and democratic norms, such as claims of discrimination, government dysfunction, and treason."

Toxicity is seemingly an inescapable part of social media [28, 76, 85, 110, 123]. Facebook estimates that between 0.14% and 0.15% of all views on their platform are of toxic comments [36]. This type of incivility, in addition to damaging online conversations, has been found to also damage civil institutions [15, 113] having dangerous real-world implications. Fink *et al.* [38] find that politically charged anti-Muslim hate speech on Facebook in Myanmar was a prominent aspect preceding the Rohingya genocide.

To prevent the spread of toxic content, various platforms have implemented and designed a variety of safeguards [1, 2, 36]. Other researchers have further performed in-depth studies on users' behavior to understand abusers and victims of abuse. For instance, Founta *et al.* [40] identify a set of network and account characteristics of abusive accounts on Twitter. Hua *et al.* [64] look at properties of the accounts that have heavily negative interactions with political candidates on Twitter. Finally, Chang *et al.*, Xia *et al.*, Zhang *et al.*, and Lambert *et al.* all look at the set of causes that make conversations unhealthy or toxic [78, 124, 126, 127].

2.6 The Interplay of Misinformation, Online Toxicity, and Political Polarization

Several works, close to our study, have attempted to understand how political ideology, online toxicity, and misinformation interact. Online toxicity, for instance, has been heavily associated with increased political polarization and the use of misinformation [25, 112]. Conversely, Rajadesingan *et al.* [93], find that political discussions in non-overtly political subreddits often lead to less toxic conversational outcomes. Cinelli *et al.* [25], show that misinformation about COVID-19 on YouTube promoted hate, toxicity, and conspiracy theories on the platform. Chen *et al.* [21], utilizing network-based analysis, find that misleading online videos often lead to increased incivility in their comments. Separately, Rains *et al.* [92] find that high political ideological extremism is a major factor in incivility and toxicity online. De Francisci Morales *et al.* [30] find, most markedly that the interaction of individuals of different political orientations increased negative conversational outcomes. Similarly, Kim *et al.*, Kwon *et al.*, and Shen *et al.* all find that exposure to these negative conversations actually increases observers' tendency to also engage in incivility [74, 77, 106]. Finally, Imhoff *et al.* [68] find that political polarization is a key aspect of people's belief in false narratives. However, despite this panoply of research, it is unclear how political ideology and toxicity interact in the presence of misinformation and across different political environments. In this work, we seek to fully understand these dynamics.

3 DATASETS & METHODS

Many previous works have investigated misinformation, toxicity, and political polarization individually; we thus rely, in several places, on previous studies when compiling our datasets. In this section, in addition to providing an overview of these datasets, we give an overview of how we calculate the political ideology of users and subreddits, how we determine the toxicity of posts and comments, and finally how we measure misinformation levels of subreddits.

3.1 Reddit Dataset

For this work, we study 18 months of Reddit comments and submissions from January 2020 to June 2021. To aggregate this data, we use Pushshift [13], a third-party API that collects and publishes monthly datasets of Reddit comments and submissions. Each comment and submission includes its timestamp, author's username, subreddit/community where the comment was posted, and the conversation thread where the comment was posted.[1] Using this data, we reconstruct the conversation threads for each user and subreddit. Throughout this work, we focus on English-language misinformation websites and thus we filter our dataset to include only English language comments (removing 400M comments) using the `whatlanggo` Go library.[2]

[1] We note that all data was collected prior to Pushshift falling outside Reddit's Terms of Service in April 2023
[2] https://github.com/abadojack/whatlanggo

3.2 Misinformation and Authentic News Dataset

To analyze how users interact with misinformation on Reddit, we first gather lists of misinformation and authentic websites (as a control). Specifically, we aggregate misinformation and authentic news domains previously gathered by Iffy News,[3] OpenSources,[4] Politifact,[5] Snopes,[6] Melissa Zimdars,[7] and Hanley *et al.* [59]. Our final list of misinformation outlets consists of 541 websites, which encompass sites like theconservativetreehouse.com and infowars.com [59]. Separately, our list of authentic news sites consists of 565 websites from across the political spectrum, including sites like cnn.com and dailywire.com.

We note, that to verify our findings, we rerun several of our initial experiments with an additional separate list of misinformation and authentic news websites (Appendix A). As our second set of misinformation websites, we utilize a set of 932 websites labeled as "questionable sources" by the website Media Bias/Fact Check.[8] Media Bias/Fact Check labels a website as a "questionable source" if it exhibits one of the following "extreme bias, consistent promotion of propaganda/conspiracies, poor or no sourcing of credible information, a complete lack of transparency and/or is fake news." This largely matches our definition of misinformation websites outlined in Section 2.1. This list of websites has been utilized throughout prior works [9, 26]. After removing websites that overlap with the first set of misinformation websites, this list contained 835 unique misinformation websites. As our second set of authentic news websites, we utilize a set of 1,885 news websites labeled as center,[9] center-left,[10] and center-right[11] by Media Bias/Fact Check. After removing duplicates from our original list of 565 websites, we arrived at a final set of 1,720 websites in our second authentic news dataset.

3.3 Misinformation Levels, Misinfo-Oriented Domains, and Mainstream-Oriented Websites

While we collect a total of 1,376 misinformation and 2,285 authentic news websites, we note that these are a subset of the *many* misinformation and news outlets on the Internet. To better approximate levels of authentic and misinformation news, we define a larger class of 157,605 domains that are *misinfo-oriented* and 667,848 that are *mainstream-oriented*.

As in prior work [59], we define *misinfo-oriented* as websites that have more connections from our set of misinformation websites than from authentic news sources (*i.e.*, the majority of a site's inward links in a domain-based graph are from our set of misinformation websites). Similarly, we define websites as *mainstream-oriented* websites that have more connections from authentic news websites than from misinformation websites. To determine which websites fall into these definitions, we utilize Common Crawl data[12]–widely considered the most complete publicly available source of web crawl data. For each misinformation and authentic news in our first dataset, we collect the set of their domain's HTML pages that were indexed by Common Crawl before August 2021. For each HTML page indexed by Common Crawl, we parse the HTML and collect hyperlinks to other pages (*i.e.*, HTML <a> tags). Using this approach, we then determine which misinformation and

[3] https://iffy.news/index
[4] https://github.com/several27/FakeNewsCorpus
[5] https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
[6] https://github.com/Aloisius/fake-news
[7] https://library.athenstech.edu/fake
[8] https://mediabiasfactcheck.com/fake-news/
[9] https://mediabiasfactcheck.com/center/
[10] https://mediabiasfactcheck.com/leftcenter/
[11] https://mediabiasfactcheck.com/right-center/
[12] https://commoncrawl.org/

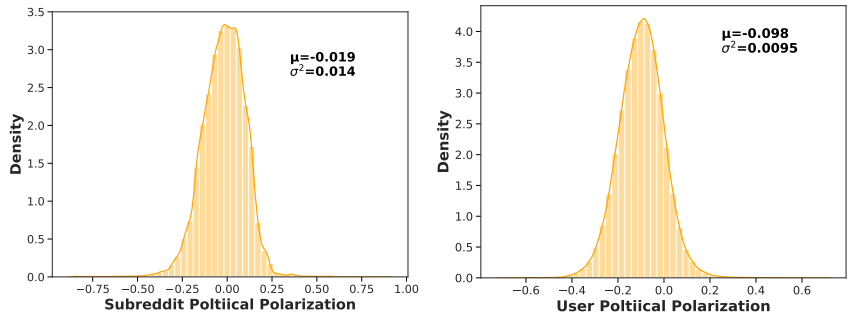


Fig. 1. Subreddit and user political polarization scores — We estimate the political polarization of users and subreddits based on the political polarization of the URLs they post. We compute these estimates for users and subreddits that have posted at least 10 URLs to get robust averages for each subreddit and user. Altogether we get approximate political polarization scores for 427K users and 46.7K subreddits.

authentic news websites have hyperlink connections with which websites on the Internet. We then calculate which websites hyperlinked by our set of misinformation and authentic news websites are *misinfo-oriented* and *mainstream-oriented*. Altogether we gather the available Common Crawl pages and scrape the HTML for 541 misinformation and 565 authentic news websites in our first URL dataset (we do not do all websites in our dataset given issues with the 100s of TBs required Common Crawl data). Websites that have been widely documented as spreading falsehood and conspiracy theories are included within this list as *misinfo-oriented* including waronfakes.com and 8kun.top [57, 58, 107]. Conversely, our list of *mainstream-oriented* websites includes reputable sources like nytimes.com and wsj.com [125].

3.4 Approximating the Political Ideology of Subreddits and Users

To approximate the political ideology of subreddits and users, we determine how often each user and subreddit respectively post conservative-leaning and liberal-leaning hyperlinks using a dataset of website partisanship scores developed by Robertson *et al.* [97]. Robertson *et al.*'s original dataset measured the partisanship of different sites based on how often they were shared by Democrats and Republicans on Twitter in late 2017. Their dataset includes partisan bias scores for 19K websites, scoring each between -1 (liberal/Democratic-leaning) and +1 (conservative/Republican-leaning). To estimate the approximate political leaning of subreddits and users, we take the average of the political partisanship scores of the hyperlinks that they posted online. For example, if a user frequently posts hyperlinks to both nytimes.com (-0.2602 Democratic/Liberal) and veteranstoday.com (+0.2994 Republican/Conservative), this would result in a political partisanship score of 0.0392. As found by Saveski *et al.* [99], utilizing the polarization of URLs posted by users was found to largely correlate (R^2 = 0.81) with users' US voting behaviors. We further note that while many of these subreddits and users may not be overtly political, their use of politically charged and biased URLs does allow us as in Saveski *et al.* [99] to approximate their political leanings.

To build a robust political polarization score for each user and subreddit, we utilize averaged scores for only users and subreddits who have posted more than 10 URLs. Furthermore, we note, that to approximate user political leanings we utilize all URLs posted by the user both in their Reddit submissions as well as their comments. In contrast, we only utilize the URLs linked in posts/submissions on subreddits when calculating a subreddit's political leaning. We make this distinction because these hyperlinks are implicitly approved by the subreddit's community and

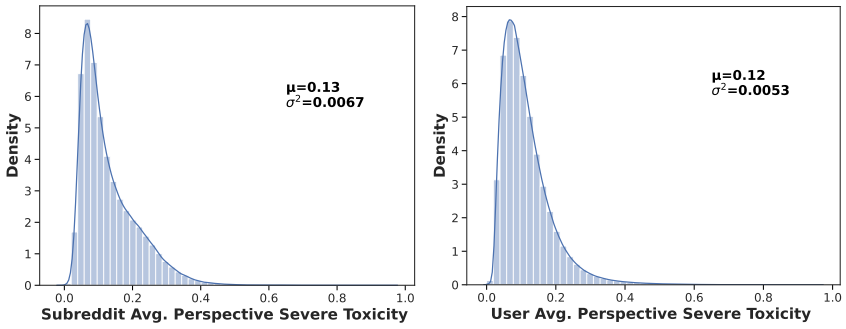


Fig. 2. Subreddit average and user average Perspective Severe Toxicity scores — We determine the toxicity norms for subreddits with at least 50 comments and users with at least 10 comments. Each user and subreddit has distinctive toxicity norms, posting toxic comments at different rates. At a threshold of 0.8, most users and the subreddit's usual comments/posts are not considered toxic or pernicious by the Perspective API SEVERE_TOXICITY classifier.

are more reflective of the political leanings of the subreddit as a whole [118]. We further remove internal Reddit hyperlinks when calculating the political leaning of users (*i.e.*, a Reddit user or subreddit hyperlinking to another page on Reddit does not affect the political leaning calculation.) Altogether, we calculate and utilize scores for 427K users and 46.7K subreddits.

As seen in Figure 1, the average political leaning of Reddit users is liberal/Democratic-leaning ($\mu = -0.0984$). This largely agrees with Pew Research polling data, which found that 47% of Reddit users identify as liberal, 39% as moderate, and 13% as conservative [12]. In contrast, we see across our measured subreddits, that the average subreddit is only slightly liberal-leaning ($\mu = 0.0186$). Agreeing with past work, this confirms that while subreddits are created by and for individuals across the political spectrum [94], the liberal/Democratic-leaning subreddits are the most popular and have the most users.

3.5 Identifying Toxic Comments and Approximating User and Subreddit Toxicity

To approximate the toxicity of Reddit users and subreddits, we utilize the Perspective API, a set of out-of-box toxicity classifiers from Google Jigsaw [2]. The Perspective API takes comments as input and returns a score of 0–1 for several classifiers. For each classifier, the closer a comment's score is to 1, the more likely the comment is pernicious or toxic. To pinpoint explicit examples of highly toxic comments, we utilize the `SEVERE_TOXICITY` classifier. The Perspective API has been utilized extensively in prior works [76, 94, 101] and we rely on the best practices outlined in past works for our study. As in Chong *et al.*, Han *et al.* [55] and other works, to consider a comment as toxic, we utilize a threshold of 0.8 [23, 78]. As found by Kumar *et al.* [75, 76], utilizing this particular classifier, while limiting recall, provides an acceptable precision for identifying toxic online content.

To calculate toxicity norms and identify toxic comments, we first determine the approximate toxicity norms for each of the 46,681 subreddits for which we have political data. When calculating toxicity norms, we further filter down to those subreddits with at least 50 comments. For determining *user* toxicity, we first identify 31.1M users within our set of 46.7K subreddits for which we have political ideology data, gathering all the comments they posted between January 2020 and June 2021 across every subreddit they posted in and retrieving their `SEVERE_TOXICITY` score with the Perspective API. We do this across *all* of these users' English-language comments in every subreddit to approximate toxicity norms for their overall behavior across Reddit. We then filter down these

users to those who have posted at least 10 comments [94]. From the returned toxicity scores, we approximate each subreddit's and user's toxicity norms by how often they post toxic content (comments with SEVERE_TOXICITY ≥ 0.8). As seen in Figure 2, while there is a wide range of online toxic behaviors, based on our strict definition of toxicity, most users and subreddits are on average non-toxic in their interactions.

3.6 Ethical Considerations

Within this work, we largely focus on identifying large-scale trends in how different subreddits interact with misinformation, levels of toxicity, and levels of political polarization. While we do calculate toxicity and polarization levels for individual users, we do not display their usernames in this work, nor do we attempt to contact them or attempt to deanonymize them. We note that all Reddit submissions and comments analyzed in this work are *still* public and *still* available through the Pushshift API [13].

4 RQ1: TOXICITY AND POLITICAL IDEOLOGY IN MISINFORMATION POSTS

In this section, we examine the intersection of toxicity levels, political ideology, and the presence of misinformation, within particular subreddit submissions. After examining the distributional differences between the users and subreddit characteristics among misinformation and authentic news submissions, we finish this section by fitting a linear model in order to understand the full degree to which each of these features collectively predicts the toxicity of user comments.

4.1 Setup

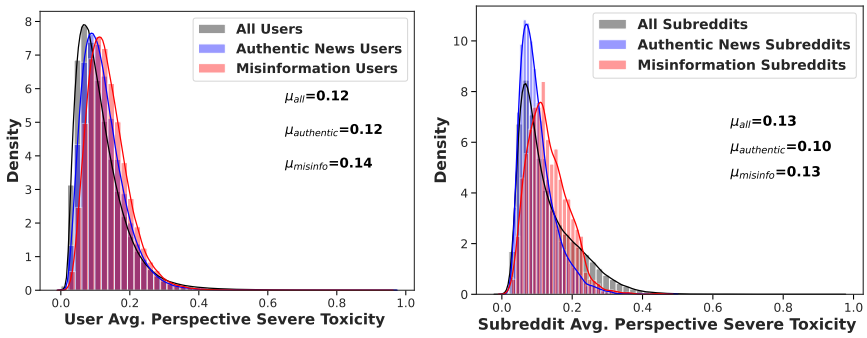
On Reddit, users can submit links to news articles as *submissions* on which users can comment. To understand the difference in levels of political ideology and toxicity associated with posts that link to misinformation websites, we compare the political ideology of users and comment toxicity in response to misinformation and authentic news URL submissions. Across all of our measured subreddits, we gather the sets of URL submissions that utilize our set of misinformation and authentic news websites. Altogether, there were 38.3K submissions in 2.2K subreddits that link to our first set of 541 misinformation websites and 227K submissions from 18.4K subreddits that link to our set of 565 authentic news websites. The difference in the magnitude of submission is likely due to the greater popularity and widespread appeal of authentic mainstream news compared with alternative, fringe websites [59]. Indeed, utilizing the Amazon Alexa Top Million list from March 1, 2021 [6], we find that 255 authentic news websites were in the top 100K websites, while only 101 misinformation websites were in the top 100K sites.

We confirm our results in this section using our second set of misinformation and mainstream websites. Altogether, from this second set of URLs, we find an additional set of 9.6K misinformation and 561K authentic news submissions. Obtaining highly similar results compared to our first set of URLs, we report this second set of results in Appendix A.

4.2 Differences in Toxicity in Response to Misinformation and Authentic News Posts

Looking at the toxicity of comments on submissions that link to misinformation sites, we see that 14.9% of submissions had at least one toxic comment and 1.80% of comments were toxic. In contrast, for our set of authentic news submissions, 13.6% of submissions had at least one toxic comment and 1.05% of the comments were toxic. We thus see an approximate 71.4% relative increase in the rate at which toxic comments are posted in response to misinformation submissions.

Higher toxicity could caused by (1) more toxic/uncivil users participating in conversations about misinformation, or (2) higher toxicity norms in the subreddits where misinformation is posted (rather than misinformation submissions simply inciting more toxic responses.) In Figure 3, we



(a) User Toxicity (b) Subreddit Toxicity

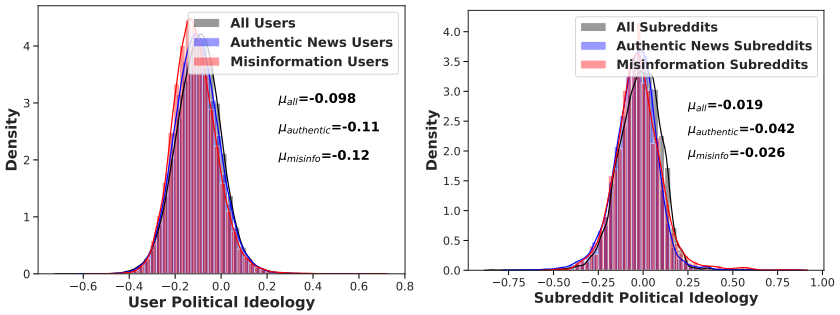
Fig. 3. Toxicity levels for users who comment under authentic news and misinformation URL Reddit submissions— Users who interact with misinformation submissions are slightly more toxic than users who interact with authentic news. Both groups are slightly more toxic than Reddit users generally. Similarly, subreddits with misinformation submissions are overall more toxic compared with authentic news subreddits and subreddits more generally.

see that users who comment on misinformation posts *are* generally slightly more toxic than their authentic news counterparts. On average 1.54% of the comments from users who respond to misinformation submissions are toxic compared with 1.22% for the corresponding group of users who comment on authentic news stories. We note that despite the proximity in the toxicity of misinformation commenters and authentic news commenters, the higher user toxicity appears stable even among users from the same subreddits. Comparing only the users who posted in subreddits where *both* mainstream and misinformation URLs were posted, we still see that the users who posted on misinformation submissions had elevated rates of toxicity (1.45% compared to 1.21%). We thus see "more toxic" users are indeed commenting more on misinformation submissions compared to authentic news submissions.[13] However, despite finding that more toxic users are indeed commenting more often on misinformation submissions, their higher rate of toxicity is not enough to explain the larger number of toxic comments on misinformation submissions. After accounting for the higher rate of user toxicity across all the URL submissions, we still see 35.5% more toxic comments than would be expected.

Other factors, besides the specific users that comment on misinformation, are predictive of a higher rate of toxicity on misinformation submissions. Examining the role of subreddits in promoting toxicity in Figure 3, we find that the toxicity norms of subreddits with misinformation submissions also correlate with higher levels of toxic comments. Altogether, for corresponding subreddits with misinformation submissions, 1.40% of comments are toxic/uncivil compared to 0.80% in authentic news submissions.

Confirming these results with our separate set of URLs (Appendix A), we thus conclude that across our examined cases, (after taking into account both the toxicity users and the toxicity norms of the corresponding subreddits), there is a heightened level of toxicity within conversations on misinformation submissions compared to authentic news submissions.

[13] We note that we perform Mann Whitney U-tests to ensure that there are indeed statistically significant differences between the rate of toxicity in misinformation and authentic news users; running these tests and finding p-values $< 10^{-12}$, we indeed conclude that both groups URL submission commenters that there are indeed higher rates of toxicity for the misinformation users.



(a) User Political Ideology

(b) Subreddit Political Ideology

Fig. 4. Political Ideology of users who comment under authentic news and misinformation Reddit submissions– There are no significant differences in political ideology between users who comment on misinformation and those that comment on authentic news. Similarly, there are no significant differences in the political ideology of subreddits where misinformation and authentic news appear.

4.3 Political Ideology among Misinformation and Authentic News Submissions

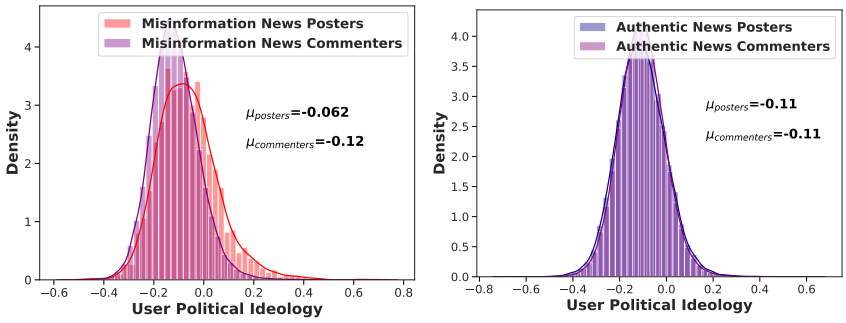
Having seen the higher levels of toxic responses to misinformation posts, we now explore the political differences between users who comment on misinformation and those who comment on authentic news. Surprisingly, we do not see dramatic differences in political ideology between the users who comment on misinformation and authentic news posts (Figure 4). For our set of misinformation URLs, we see a slight leftward tilt in the average commenter (-0.12 vs. -0.11). Similarly, we see little difference in the political ideology of subreddits where misinformation and authentic news submissions appear. Broadly, both misinformation and authentic news appear within subreddits across the political spectrum and are commented on by users across the political spectrum.

However, while misinformation appears in subreddits across the political spectrum, the users who post misinformation submissions have a rightward tilt compared to the users who comment on misinformation posts. As seen in Figure 5, users who post misinformation are, on the whole, more conservative than their corresponding more liberal commenters. In contrast, as seen in Figure 5 authentic news posters and commenters share nearly the same distribution. Altogether, we observe (especially in contrast to authentic news submissions), that a politically different set of users post misinformation news compared to those that comment on it. Thus while we do not observe that the political ideology levels of users who comment on misinformation are substantially different from commenters on authentic users, we do observe that they *are* different from posters of misinformation content.

We again confirm these findings utilizing our second set of misinformation and authentic news domains, reporting the results in Appendix A.

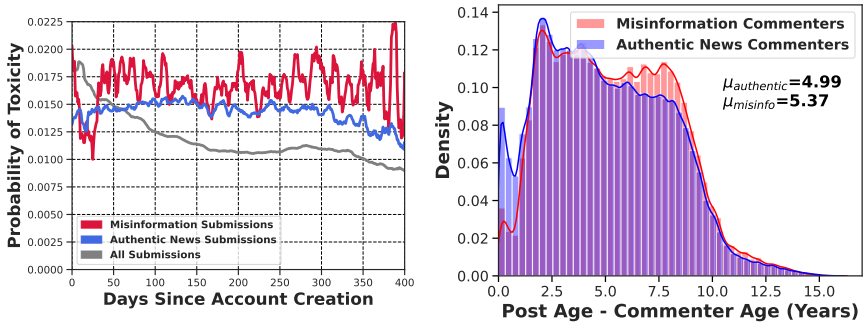
4.4 Account Age and Type among Misinformation and Authentic News Submissions

As suggested in prior work [95], the accounts commenting on misinformation submissions could be newer or throw-away accounts (which tend to be more toxic than older accounts), increasing overall levels of toxicity and potentially confounding our analysis. Looking across all of Reddit, as seen in Figure 6a, we do indeed see that newer accounts *are* more toxic than older accounts. However, we also see that misinformation commenting accounts tend to be older than mainstream



(a) Misinformation (b) Authentic News

Fig. 5. Political Ideology of posters and commenters of misinformation and authentic news— There is a noticeable rightward tilt in users who post misinformation compared to those who comment on misinformation. Unlike for misinformation posts, the posters and the commenters on authentic news share similar distributions of political ideology.



(a) Creation Date vs Toxicity (b) Age of Commenters

Fig. 6. Newer Accounts are more likely to be toxic—Across Reddit, the longer that a user has been on Reddit the less likely they are to post a toxic comment. For misinformation and authentic news submissions, this relationship, however, is less pronounced. Furthermore, on average, users who comment on misinformation tend to have slightly older accounts than authentic news commenter accounts.

commenter accounts (Figure 6b) *and* the relationship between age and account toxicity is less pronounced for misinformation and authentic news commenters (Figure 6a).

Another potential confounder that can cloud our analysis is the type of accounts that post under mainstream and misinformation accounts. From Table 1, we observe that both Mod/Admin accounts[14] (accounts that oversee and regulate discussion on subreddits) increased toxicity compared to non-moderator users across all news-related submissions, largely in contrast to submissions as a whole. We note, despite this, increased overall toxicity in news-related submissions, in general, when admins/moderators respond to toxic comments within subreddits, they typically do not respond in a toxic manner (Table 2). Furthermore, we find that admins/moderators are more likely to respond to toxic comments in news-related subreddits compared. While not responsible for

[14]We note that moderator and admin Reddit accounts make up a relatively small percentage of these toxic interactions; 4.1% of toxic comments on misinformation submissions and 2.9% of toxic comments on authentic news submissions. 98.2% of all toxic interactions do not involve a moderator/admin. Nearly all activity, including toxic activity, is by non-mod/admin users.

Misinformation Submissions	
Mod/Admin Toxicity	2.20% ± 0.20%
Non-Mod/Admin Toxicity	1.75% ± 0.02%
Authentic News Submissions	
Mod/Admin Toxicity	1.17% ± 0.04%
Non-Mod/Admin Toxicity	0.76% ± 0.005%
All Submissions	
Mod/Admin Toxicity	0.46% ± 0.000%
Non-Mod/Admin Toxicity	1.36% ± 0.000%

Table 1. Percent of Toxic Comments for Moderator/Admin and Non-Moderator/Admin Users on Misinformation, Authentic News, and All Submissions with 95% Normal Confidence Intervals.

	Misinformation Submission Responses	**Mainstream Submissions Responses**	**All Submissions Responses**
Mod/Admin Responding Toxicly to Toxicity	0.048%	0.030%	0.036%
Mod/Admin Responding Nontoxicly to Toxicity	1.02%	0.80%	0.48%

Table 2. Moderators and admins have similar roles across different submissions. Admins/moderators tend to respond more to toxic comments with nontoxic comments.

the difference in toxicity between misinformation and authentic news submissions, given the low percentage of moderator/admin comments overall, we thus see that admin status as well as the user-creation data is associated with differences in the relative toxicities of Reddit accounts. Having observed that both account age and account type are potential confounders, in our next section, we include these factors when collectively analyzing the features predictive of a Reddit submission's toxicity.

4.5 Intersection of Misinformation, News Media, Toxicity, and Political Ideology

Finally, having analyzed the potential role of user-level and subreddit-level factors in the toxicity of comments posted on misinformation and mainstream Reddit submissions, we determine whether these characteristics correlate with each subreddit's similarity to misinformation and authentic news. To do this we rely on our list of *misinfo-oriented* and *mainstream-oriented* websites. For each subreddit in our dataset, we compute their *misinformation similarity* and their *mainstream similarity* based on the percentage of each subreddit's URL submissions that come from websites that are *misinfo-oriented* and *mainstream-oriented*. This measurement essentially determines the approximate percentage of submissions within each of our subreddits that is misinformation-oriented/related and the percentage that is mainstream-oriented/related. Lastly, we note, that to confirm the results of this section and to fully understand the role of misinformation similarity and authentic news similarity in submissions' toxicity, we fit a linear model to understand how each of the features previously considered, predicts the average toxicity of the comments of individual Reddit submissions.

As seen in Figure 7, across our 46.7K considered subreddits, we observe that as subreddits become more similar to misinformation websites and hyperlink to more *misinfo-oriented* domains, their overall level of toxicity increases. This largely matches our observation in Section 4.2 that misinformation submissions are in general more toxic than authentic news submissions. We, however, as further seen in Figure 7, misinformation similarity is not heavily correlated with political ideology. Conversely, we do not see a significant correlation between subreddits' mainstream similarity and their overall level of toxicity. This reinforces our results that mainstream news

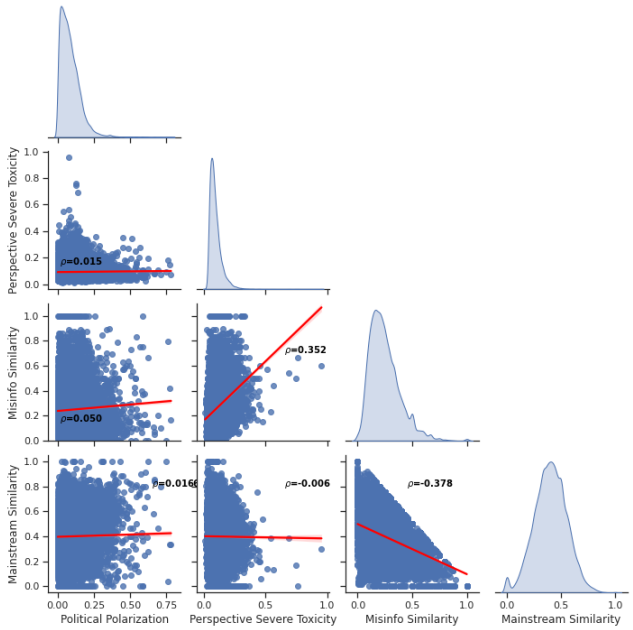


Fig. 7. Misinformation, toxicity, and political ideology interactions—As subreddits increase in misinformation similarity, they become more toxic. However, there is not a large correlation between misinformation similarity and the political ideology of subreddits. Similarly, we do not see any correlation between political ideology levels and mainstream similarity; nor do we see any correlation with toxicity levels.

Adjusted R-squared: 0.282	**Coefficient**
Intercept	-0.0373^{***}
Is a Misinformation Submission	0.0028^{**}
(Abs) Subreddit Political Ideology	$+3.73 \times 10^{-5}$
Average Reddit Thread (Submission Date-User Creation Date)	-0.0004^{***}
Moderator/Admin Involvement	$+0.0010^{*}$
(Abs) Average User Ideology of Submission Commenters	-0.00664^{*}
Average User Toxicity of Submission Commenters	$+1.184^{***}$
Reddit Thread User ideology (Std)	$+0.0144^{*}$
Subreddit Mainstream Similarity	$+0.0038$
Subreddit Misinformation Similarity	$+0.0280^{***}$

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

Table 3. Model of the toxicity of the comments in Reddit submission. The average toxicity of the users involved in a given Reddit thread is the primary determinant of the toxicity within that Reddit threat. As a given subreddit shares more misinformation-related URLs, the toxicities of the corresponding subreddit's threads also increase. Note that we do not include subreddit toxicity as a feature in this analysis as each Reddit thread's toxicity was directly utilized to calculate this value.

similarity is not correlated with toxicity and that neither misinformation similarity nor authentic news similarity is highly correlated with political ideology(Section 4.2 and Section 4.3).

Finally, to confirm these results, we fit a linear regression of the average toxicity of the comments of individual submissions against the subreddit and user-level variables explored within this and the previous three sections. *As seen in Table 3, while still accounting for (1) the political ideology of the subreddit, (2) the average relative age of the accounts posting, (3) moderator and admin involvement within the submission thread, (4) the toxicity of individual users, (5) the standard deviation of political ideologies involved in the Reddit thread,* we find that misinformation levels within a given subreddit have a statistically significant and positive effect on levels of toxicity, while levels of mainstream similarity do not. In addition to observing that a subreddit's misinformation similarity is positively associated with a given Reddit submission in that subreddit having higher levels of toxicity, we

further observe that **all** of the associations previously found in this section continue to have their corresponding associations with toxicity. Namely, we find that moderator/admin involvement, individual user toxicity, the breadth of political ideologies in a given Reddit thread, and a subreddit level of misinformation similarity are all positively associated with increased toxicities. We similarly find that subreddits' mainstream similarity, subreddits' political ideology level, and commenters' political ideology level are not associated with increased toxicity.

4.6 Summary

In this section, we found that misinformation on Reddit largely is correlated with and predictive of higher amounts of toxicity. Most markedly, we observed that the comments under misinformation submissions are posted at a rate 71.4% higher than the comments under authentic news submissions. Further, while we do observe a dichotomy in the political ideology of users who post misinformation and those who comment on misinformation, somewhat surprisingly, we find that misinformation appears across different political environments, with it not being concentrated just in the political extremes. Lastly, looking at how different levels of misinformation correlate with toxicity, we find the more *misinfo-oriented* submissions a given subreddit has, the more toxic/uncivil it is likely to be. We confirm these relationships by fitting a linear regression of all the examined features against the toxicity of each of the misinformation and mainstream submissions.

5 RQ2: MISINFORMATION AND POLARIZED TOXIC INTERACTIONS

In the previous section, we showed that comments on misinformation submissions are 71.4% more toxic than those on authentic news submissions. Furthermore, there appears to be a difference in the political orientation of the users who post misinformation and those who comment on it. Given this difference and the higher toxicity comments on misinformation submissions, we now turn to understand whether *political differences* are correlated with increased toxic individual interactions within Reddit misinformation submissions.

5.1 Setup

To understand how *political differences* correlate with toxicity, we reconstruct the conversational dyads that exist underneath each Reddit submission. Reddit comments are similar to conversational threads; if a user responds to a given comment, their reply will appear underneath the comment. For each submission in our Pushshift [13] dataset, we determine using the thread information whether the commenter posted a response directly to another commenter. This enables us to reconstruct conversational dyads between individual Reddit users. Then, using the approach outlined in Section 3.1, we determine the polarization and average toxicity of the users in our conversational dyads. From these averages, we label users as right-leaning (positive political ideology score) or left-leaning (negative political ideology score).

Looking at each conversational dyad, we determine if each comment is toxic using the Perspective API (as outlined in Section 3.1). For a comparison of how conversations differ between misinformation and authentic news comments, we finally separate the set of conversational dyads that appear under misinformation versus authentic news submissions. We lastly note, having confirmed our initial findings using our dual set of URLs, we, for the rest of this work, combine these two URL lists (*i.e.*, we consider the conversational dyads under submissions that utilize our full set of 1,372 misinformation domains and 2,285 authentic news domains.)

5.2 Interactions within Misinformation and Authentic News Environments

We find a high degree of affective polarization across conversational dyads: 81.7% of interactions are between users of the same political orientation (*i.e.*, liberal–liberal, conservative–conservative).

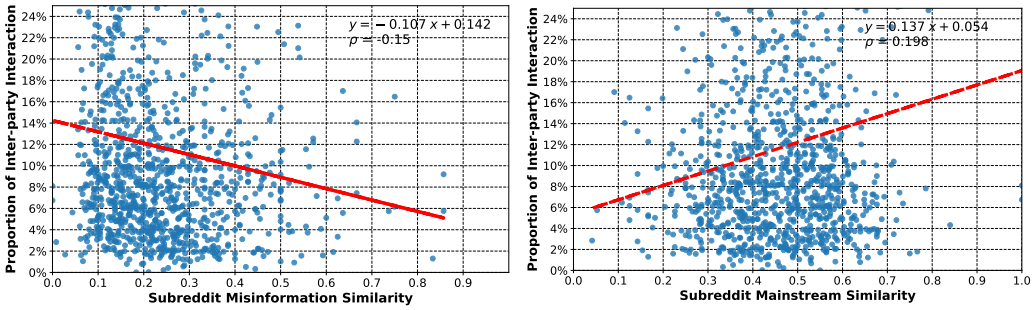


Fig. 8. Proportions of inter-party interactions—As subreddits link to more *misinfo-oriented* websites, as a percentage, there are fewer interactions between conservative and liberal users. In contrast, there is a slight correlation between hyperlinking to *mainstream-oriented* websites and more inter-party interactions.

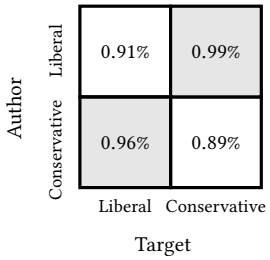
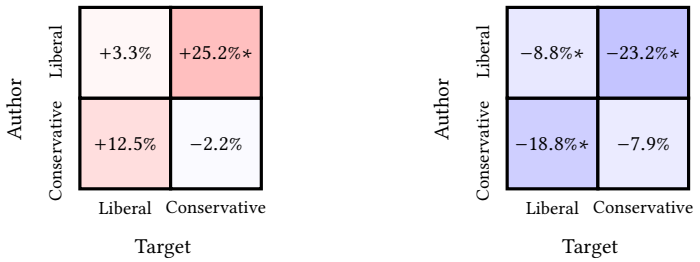


Fig. 9. Percentage of interactions that are toxic/uncivil for authors and targets of different political leanings. Across all 46K considered subreddits, there is a slight heterophily for users to reply in a toxic manner to members with a tilt towards the political group.

For conversations under authentic news and misinformation submissions, this changes to 81.3% and 85.3% respectively. Investigating subreddits with at least 50 toxic conversational dyads, we further find that as subreddits hyperlink to more *misinfo-oriented* websites, conversations become more insular ($\rho = -0.150$). In contrast, as subreddits hyperlink to more *mainstream-oriented* websites, there is a slight increase in inter-party conversations ($\rho = 0.198$). We show these relationships in Figure 8. Together, we see that posts to misinformation sites are correlated with heightened intra-party conversations, potentially creating more insular communities, while authentic news is associated with a very slight increase in inter-party political conversations.

With regards to interactions associated with misinformation posts, we observe a similar effect for toxic comments. As seen in Figure 9, across all conversations[15], we see a slight increase in toxic replies between users of different political ideologies. We calculate an odds ratio of 1.17 for users that reply in a toxic manner to users of a different political leaning compared with users of the same political leaning. Comparing the set of toxic conversational dyads under misinformation submissions, we see even greater animosity between users of different political affiliations. Compared with the baseline across all conversations, we observe a 25.2% relative increase in the percentage of liberal to conservative toxic comments and a 12.5% relative increase in the percentage of conservative to liberal toxic comments.[16] In contrast, compared with the baseline across all conversations, for authentic news submissions, we see a 23.2% relative decrease in liberal to conservative toxic comments and an 18.8% drop in the percentage of conservative to liberal toxic

[15] Across all conversational dyads, in 54.6% of the dyads, only the original commenters were toxic, in 41.0% of the dyads only the responders were toxic, and in 4.4% of dyads, both the original commenter and the responder were toxic.
[16] Across our all authentic news dyads, in 62.3% of the dyads only the original commenters were toxic, in 34.5% of the dyads only the responders were toxic, and in 3.2% of dyads, both the original commenter and the responder were toxic.



(a) Misinformation Submission comments (b) Authentic News Submission comments

Fig. 10. Percentage increases in interactions that are toxic in misinformation and authentic news submissions for conservative and liberal authors against conservative and liberal targets compared against the baseline of all interactions (Figure 9). We ensure that the respective shifts in percentage increases and decreases are significant by performing t-tests. Values that have p-values≈0 are starred. All other values were found to be non-significant (*i.e.*, p-values > 0.00625, [α=0.05/8 after Bonferroni correction.])

Misinformation Interactions	**Coefficient**
Intercept	-8.512***
Absolute User Ideology	0.434
User Ideology Differences	-0.893*
User Toxicity	12.410***
User Moderator/Admin Status	-0.3067
Relative User Age (years)	0.0028*
Reciprocity	4.573 ***
Shared Subreddits	0.00063
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Mainstream/Authentic News Interactions	**Coefficient**
Intercept	-8.711***
Absolute User Ideology	0.882*
User Ideology Differences	-1.278***
User Toxicity	9.003***
User Moderator/Admin Status	-0.4660
Relative User Age (years)	-0.0049
Reciprocity	3.569***
Shared Subreddits	0.0004
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 4. Toxic Misinformation and Authentic News Submission Interactions. As confirmed in our ERGM, differences in the political orientation of users are predictive of increased incivility and toxicity with users of differing political orientations more likely to engage in toxic interactions within misinformation submissions than on mainstream submissions. Similarly, the higher each user's toxicity norm, the more they are likely to target other users with toxic comments.

comments (Figure 10).[17] This appears to indicate that while in misinformation-laced conversations, users are more likely to respond in a toxic manner to users of a different political orientation, users in authentic news-centered conversations are less likely. To confirm, we calculated the odds ratio: 1.64 for misinformation toxic comments and 0.87 for mainstream toxic comments when comparing the percentages of politically inter-party toxic comments to politically intra-party comments.

Overall, we find that on average, when responding to misinformation posts, users are slightly more likely to respond toxically to users of the opposite political ideology compared to all submissions on Reddit. This difference, in part, helps to explain the higher levels of toxicity observed within misinformation submissions in Section 4.2 given the political differences between misinformation posters and misinformation comments.

5.3 Modeling Toxic Interactions Between Users Responding to Misinformation Posts

To more concretely show that users of different political stripes in misinformation-laced conversations are more likely to reply in a toxic manner to each other, we fit our network data of toxic interactions to an exponential random graph model. An Exponential Random Graph Model (ERGM) is a form of modeling that predicts connections (*e.g.*, toxic interactions) between nodes (users) in a

[17] Across our all authentic news dyads, in 60.1% of the original commenters were toxic, in 36.6% of the dyads only the responders were toxic, and in 3.2% of dyads, both the original commenter and the responder were toxic.

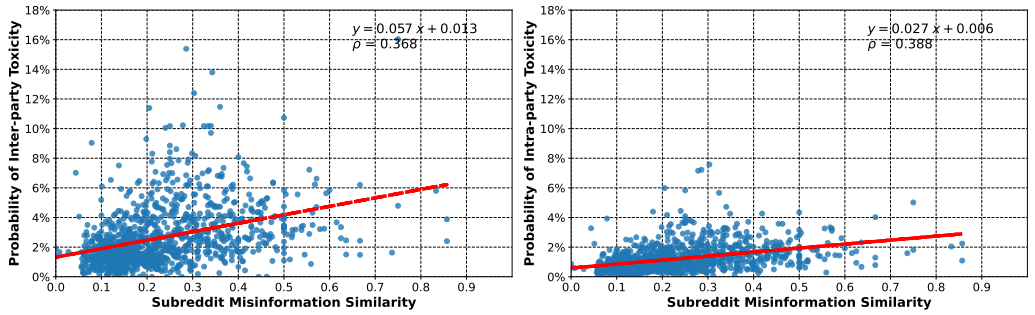


Fig. 11. Subreddit misinformation similarity vs. probability of toxic interactions between users of different and same political orientation— While for both inter-political and intra-political interactions, as misinformation similarity in a subreddit increases, the probability of a toxic interaction increases, for inter-political interactions the rate of increase is nearly double.

given network [67]. ERGM models assume that connections are determined by a random variable p^* that is dependent on input variables. As in Chen *et al.* [21] and Peng *et al.* [87], we utilize this modeling as it does not assume that its data input is independent; given that, we want to model the interactions of polarization, toxicity, this relaxed restriction is key (we have already seen that they are largely not independent) [67, 114]. Utilizing this framework, we model the probability of toxic interactions between a given author and target[18] within misinformation submissions as a function of *(1) their percentage of toxic comments, (2) their political polarization, (3) the difference in the author and target's political polarization, (4) whether they are a moderator or admin in a subreddit, (5) the relative age of the target and author, (6) the reciprocity between the author and target (i.e., if the author and target both had a toxic comment aimed at each other), and finally, (7) the number of subreddits that they share.*

Fitting our ERGM to both misinformation and authentic news conversational dyads, we see that for both cases, neither account age nor account admin/moderator status have any significant effect on interactions. Indeed for misinformation interactions, we see that as accounts become *older* relative to when they post, the more likely they are to engage in toxic interactions. From Table 4, for authentic news interactions, we further observe that the more politically polarized a user is (on either side of the political spectrum) the more likely they are to engage in toxic behavior. For all news interactions, we find that (1) that the more toxic a user, the more likely they are to engage in toxic interactions, and (2) that users are more likely to respond in a toxic manner to users who engage with them in a toxic manner (reciprocity). However, most importantly, we find that while most toxic interactions occur among users that are politically similar to each other, compared to authentic news interactions, users under misinformation submissions are *more* likely to send toxic comments to users of different political ideologies than users under mainstream submissions (-0.893 vs -1.278).

We thus have seen that not only do misinformation submissions have more insular conversations, with 85.3% of conversational dyads between users of the same political orientation (compared to 81.3% of conversations under all Reddit submissions) but also that users become more hostile to users of the opposing political orientation compared with users who post under authentic news submissions.

[18]Similar to past research on online abuse, we avoid the term "victim" to not disempower people facing abuse [75, 110].

Adjusted R-squared: 0.289	Coefficient
Intercept	0.00585***
Subreddit Misinfo Similarity	0.0270***
Type of Interaction (Intra vs Inter-Party)	0.00739***
Misinfo Similarity*Inter-Party	0.0302***

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

Table 5. Moderation Analysis on Different Types of Interactions: Fit of the probability of toxic comments in subreddits against levels of misinformation-oriented hyperlinks and the type of interactions (inter-party vs. intra-party).

5.4 Misinformation Similarity and Increased Rates of Inter-Political Toxicity

Having confirmed that users commenting under misinformation submission are more likely to engage in negative interactions with users of different political orientations, we next determine if the overall levels of misinformation within a subreddit lead to increased inter-party toxic interactions as a whole. Namely, as misinformation news sourced article levels in a subreddit as a whole increase, does the probability of negative interactions between users of different political orientations increase? We thus plot the percentage of misinformation-oriented hyperlinks within a given subreddit against the probability of toxic interaction between members of the two political orientations (Figure 11).

As seen in Figure 11, considering subreddits with at least 50 toxic conversational dyads, we see that as subreddits have more *misinformation-oriented* submissions/increased misinformation similarity, the percentage of toxic conversations dyads between users of different political leanings increases (ρ = 0.368). While we similarly see that intra-political toxicity as a function of the amount of *misinformation-oriented* hyperlinks/misinformation similarity also increases with a similar correlation ρ = 0.388, the rate at which misinformation induces inter-political toxicity is nearly 2.1 times that of intra-political toxicity (0.057 slope vs. 0.027 slope). Performing a moderation analysis by fitting a linear regression on misinformation similarity vs. probability of toxic comments with the type of interaction (intra-party vs. inter-party) as the moderation term, we indeed see the inter-party toxic increases at a faster rate than intra-party interactions (Table 5). This indicates that *misinformation-oriented* subreddits are on the whole more toxic but that they increase inter-party toxicity at a faster rate than intra-party toxicity.

Performing the same analysis and comparing against subreddit mainstream similarity, we do not see a similar relationship. As seen in Figure 12, the relationship between inter-political and intra-political toxicity rates and similarity to mainstream sources is largely flat. After fitting our linear regression and performing the same moderation analysis, in Table 6 we find, as in Section 5.2, that mainstream similarity correlates with a decreased rate of inter-party toxicity (-0.00397).

5.5 Summary

In this section, we showed that posts to misinformation outlets not only promote higher levels of toxicity but are also correlated with increased inter-political incivility. Fitting an ERGM to our toxic conversational dyads posted in response to misinformation stories, we find that political differences (along with reciprocity and each user's toxicity) drive more toxic interactions. Finally, examining how misinformation promotes toxicity among users, we find that across our considered subreddits, misinformation drives inter-political incivility at 2.1 times the rate of intra-political toxicity.

6 RQ3: ENGAGEMENT WITH MISINFORMATION AND AUTHENTIC NEWS

Having shown how misinformation is correlated with more toxic and politically insular environments on Reddit, we now determine these factors' role in user engagement with misinformation and

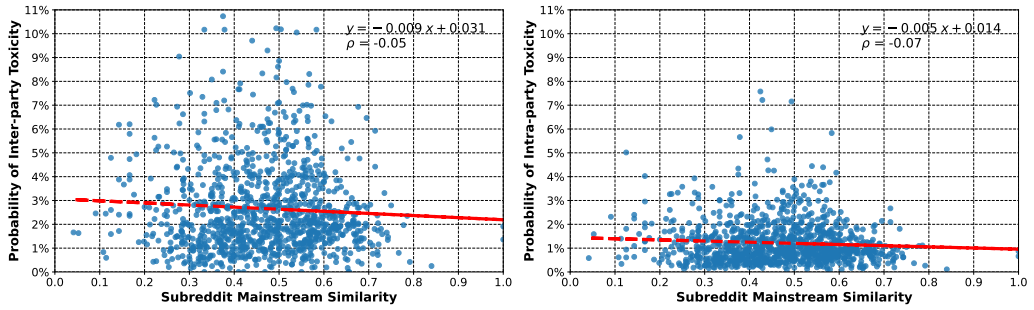


Fig. 12. Subreddit mainstream similarity vs. probability of toxic interactions between users of different and same political orientation— For both inter-political and intra-political interactions, as mainstream similarity in a subreddit increases the probability of inter- and intra-political toxicity is largely flat.

Adjusted R-squared: 0.1778	**Coefficient**
Intercept	0.0144***
Subreddit Mainstream Similarity	-0.00487
Type of Interaction (Intra vs Inter-Party)	0.0163***
Mainstream Similarity*Inter-Party Type	-0.00397**

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

Table 6. Moderation Analysis on Different Types of Interactions: Fit of the probability of toxic comments in subreddits against levels of mainstream-oriented hyperlinks and the type of interactions (inter-party vs. intra-party).

authentic news. While it is clear that misinformation drives higher percentages of conflict between users of different political beliefs, what community-level and user-level factors predict increased interaction with misinformation and thus this strife? Namely, having seen that misinformation is associated with more toxic and politically uncivil environments, are these environments also associated with more engagement with misinformation? Do users comment on and engage more with misinformation in toxic and politically insular environments?

6.1 Setup

To measure user engagement with misinformation and authentic news submissions, we utilize the number of comments that each submission receives.[19] To properly model the number of comments, we remove comments from Reddit "auto moderator" accounts (often subreddits have auto moderators that automatically comment on submissions). Finally, to model the number of comments on submissions, we utilize a zero-inflated negative binomial regression [96]. Within our regression, each observation data point represents a single submission and the number of posted comments. We utilize a zero-inflated negative binomial regression as it appropriately models our set of count data. Unlike a Poisson model, which is often utilized to model count data, negative binomial regressions do not make the strong assumption that the mean of the data is equal to the variance [83]. (Some submissions garner thousands of comments while others garner none.)

[19]We utilize the number of comments rather than the number of upvotes/downvotes, due to the unreliability of Pushshift's data for this particular characteristic. While Pushshift often can acquire most submissions and comments, it often fails to keep up-to-date information about the number of votes a given submission receives [13]. This is largely due to the high rate at which submission upvotes/downvotes change. We thus use the more stable and reliable "number of comments" number to determine user engagement with a given submission.

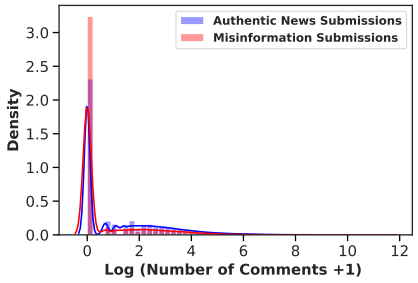


Fig. 13. Log of the number of comments on misinformation and authentic news Reddit submissions— A large majority of submissions do not receive comments.

We further utilize the zero-inflated version of this regression given the heavy preponderance of submissions that do not receive any comments. After removing comments from auto moderators, as depicted in Figure 13, 54.5% of submissions within our dataset did not receive any comments. A normal negative binomial model would be unable the correctly model this behavior.

We finally note that zero-inflated negative binomial regressions return two sets of coefficients. One set of coefficients, the zero-inflated coefficients, estimated using logistic regression, reports the probability that the given submission would receive 0 comments as a function of the covariates. Positive coefficients for these zero-inflated coefficients indicate that increases in the predictor variable make the submissions receiving 0 comments more likely. Thus the more negative a coefficient, the more the given covariate correlates with inducing at least 1 comment. The second set of coefficients, the negative binomial coefficients, model the number of comments as a function of the covariates. For these coefficients, positive coefficients indicate that the larger the corresponding covariate, the more comments that submission was likely to have received. We thus, in our analysis, can understand how different covariates affect the probability that a given submission will receive *any* comments *and* how these same covariates affect the number of comments received.

For data, we model the number of garnered comments for both our set of 47,822 misinformation submissions and 787,603 authentic news submissions. *As factors influencing the number of comments, we utilize (1) the user's admin/moderator status, (2) the relative age of the account that posted the submission, (3) the submitter's political ideology, (4) the subreddit's polarization, (5) the toxicity norm of the subreddit, (6) the submitter's toxicity norm, and (7) the average number of comments with the subreddit the submission was posted in.*

6.2 Results

Before engaging in a thorough analysis of the fits of our zero-inflated negative binomials, we first spot-check our results: we ensure that the higher the average number of comments in a given subreddit, the more likely a submission is to get comments *and* that this average correlates with more comments on submissions. In other words, we check that submissions in subreddits where users comment more also see more comments. As seen in both Tables 7 and 8, for both misinformation and authentic news Reddit submissions, as the average number of comments in a subreddit increases, (1) the more likely a submission is to receive comments and (2) the more comments it is likely to receive. Having observed this behavior, we now examine the rest of the covariates within our fits (Tables 7 and 8).

User Admin/Moderator Status. For both misinformation and authentic news submissions, we observe when user/moderator accounts post, they are less likely to garner comments. However, if a moderator/admin-submission *does* receive comments, the post is more likely to receive more comments than a non-moderator/admin post.

Account Age. For misinformation and authentic news submissions, we do not find a significant coefficient for the age of a submitting account and whether the submission receives *any* comments. However, we do find that as account age increases, both misinformation and authentic news submissions are more likely to receive comments. This may indicate that accounts with more history may attract more engagement with their posts due to their reputation or knowledge.

User Political Ideology. For both misinformation and authentic news submissions, the political ideology of the posting user has similar effects. Namely, for both authentic news and misinformation submissions, we see that as the submission's submitter becomes more politically ideological (*i.e.*, moves to the political extremes on the political left or right), the more likely their posts are to receive comments. With zero-inflated coefficients of -7.24 for misinformation submissions and -2.74 for authentic news submissions, we see that this is particularly true for misinformation submissions. This result is in line with prior work that has shown that highly ideological users are likely to provoke and garner comments on social media platforms [63, 73].

Number of Comments on Misinformation News Websites Submissions		
	Zero Inflated negative coefficient = more likely to get comments	Negative Binomial positive coefficient = more comments
Intercept	4.557***	3.442***
User Moderator/Admin Status	1.825***	2.513***
Submission Date- User Creation Date	0.021	0.045***
Absolute User Ideology	-5.635***	-2.584 ***
Absolute Subreddit Ideology	2.123***	-3.565 ***
Subreddit Toxicity	-2.439*	-2.394***
User Toxicity	-11.519***	-6.206**
Average # Subreddit Comments	-5.902***	0.884***
	$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 7. Fit of our zero-inflated negative binomial regression on the number of comments on our set of misinformation URL submissions across different subreddits.

Number of Comments on Authentic News Websites Submissions		
	Zero Inflated negative coefficient = more likely to get comments	Negative Binomial positive coefficient = more comments
Intercept	3.471***	1.694***
User Moderator/Admin Status	0.548***	1.480***
Submission Date-User Creation Date	0.024	0.082***
Absolute User Ideology	-3.066***	-1.212***
Absolute Subreddit Ideology	5.685***	-1.025***
Subreddit Toxicity	6.019***	8.736***
User Toxicity	-13.966***	-3.534***
Average # Subreddit Comments	-6.455***	0.747***
	$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 8. Fit of our zero-inflated negative binomial regression on the number of comments on our set of authentic news URL submissions across different subreddits.

However, despite highly ideological users being able to attract at least one comment, we observe that for both authentic news and misinformation submissions, as the posting user becomes more politically ideological, the fewer comments their post is likely to receive. This appears to indicate that in the case of misinformation and authentic news submission, Reddit users are perhaps being "turned off" and are engaging less with highly ideological users [61] compared to more politically neutral users.

Subreddit Ideology. We find that for both authentic news and misinformation submissions, the more politically ideological a subreddit is, the less likely anyone is to comment. This is particularly true for authentic news submissions (5.685 vs. 2.123). This may indicate that news posts, and, in particular, authentic news submissions, do not ordinarily gain traction on highly polarized subreddits. Rather, as documented by Wang *et al.* [118] subreddits like these often ignore more trustworthy sources.

In contrast, for both misinformation and authentic news submissions, we find that as polarization goes up, the more comments given submissions are likely to garner. This is particularly true for misinformation submissions (-3.565 vs. -1.025). This reflects that *when* authentic news and misinformation submissions are noticed, the more polarized the environment, the more users comment on and engage with submissions [73].

Subreddit Toxicity. Looking at the subreddit toxicity coefficient in predicting whether a submission receives comments, we see a marked difference between authentic news submissions and misinformation submissions. We see, notably, for misinformation submissions, the more toxic a subreddit is, the more likely the submission is to get comments. In contrast, for authentic news submissions, the more toxic the subreddit, the more likely the submission is to not get any comments at all. As a result, in more toxic environments, it appears that these types of submissions may be ignored. In contrast, oftentimes misinformation websites often post inflammatory articles designed to engender angst in their readership. For example, with regards to the COVID-19 pandemic, the misinformation website battle.news [107] recently published a report entitled *"Wake Up! Even The Masks Made You Sick"!*[20]

However, we further find, for authentic news submissions, that as subreddit toxicity increases, the more comments submissions are likely to garner. In contrast for misinformation submissions, the more toxic the subreddit, the fewer comments the submission is likely to garner. This reflects that *when* authentic news submissions are posted, the more toxic the environment the more users seem to comment and engage with the submissions. In contrast, when misinformation is noticed in toxic environments, this appears to not draw extensive interactions; rather the less toxic the environment, the more likely that people are to comment on the misinformation post. We thus see that authentic news submissions are more often ignored in toxic subreddits when compared to misinformation and simultaneously that as communities get more toxic, they tend to comment more on authentic news and less on misinformation submissions.

User Toxicity. Finally, looking at submitting user's toxicity, we see similar behaviors for both authentic news and misinformation submissions. Most notably, as the submitting users become more toxic, for both misinformation and authentic news submission, they are more likely they are to provoke at least one comment. We thus see that user toxicity, like political polarization, is a means by which to gain engagement generally. However, again in both cases, we see that while user toxicity often provokes at least one person to react, we see that this toxicity, often does not lead to more comments on the whole. As found in prior work, toxic users, while often sparking retorts as other users become enraged, also create unhealthy, short, and otherwise bad conversational

[20] https://web.archive.org/web/20220801105629/https://battleplan.news/watch?id=62cf06f3c0f117796a9553b7

outcomes [76, 101]. This result largely matches our definition of individual toxicity as comments that *are likely to make one leave the discussion* from Section 2.1.

6.3 Summary

In this section, we find that user toxicity, subreddit toxicity, account age, and account type, all play similar roles in the number of comments a particular submission receives across both misinformation and authentic news article submissions., However, most notably, we find that subreddit toxicity has a markedly different role in the number of comments that misinformation news website article submissions receive versus authentic news article submissions. We observe that as subreddit toxicity norms become higher, misinformation news articles are more likely to get at least one comment, while authentic news articles are less likely to get any comments. Conversely, as subreddits' toxicity norms become higher, misinformation news articles get fewer comments while authentic news articles garner more comments.

7 LIMITATIONS

In this work, we used a large-scale approach to understand the role of misinformation in insular and toxic communities online. We outline the limitations of our approach in this section.

Misinformation. One of the limitations of our approach is our use of hyperlinks to determine the presence of misinformation and estimate political polarization levels. Our approach relies on the presence of particular US-based domains on given subreddits and largely only measures US-centric misinformation and polarization. As a result, we are largely unable to extrapolate our results to non-English subreddits and non-US-based political environments. However, we note, that while our work centers on US-based political environments, as found in prior works, highly political environments across different cultures often utilize misinformation and often share many of the same characteristics as US ones [58, 68]. We leave the full investigation of this phenomenon on Reddit to future work. We similarly note that while we utilize hyperlinks to estimate polarization and this approach has been used before [100], we are unable to take into account instances when users or subreddits link to particular mainstream or conservative or liberal leaning articles to merely ridicule them (*i.e.,* these instances would moderate users' and subreddits' political ideology). Furthermore, as we examined much of Reddit using our approach, we were unable to take a comment-by-comment-based approach to understand the levels of misinformation. As a result our approach inevitably missed out on some of the subtleties of the misinformation in different subreddits. However, as found in several past works [58, 62, 102, 117], examining misinformation from a domain-based perspective enables researchers to track readily-identifiable questionable information across different platforms and is a reliable way of understanding the presence of misinformation in large communities or websites (*e.g.*, subreddits).

Measuring Toxicity. Another limitation of our approach, given our use of hyperlinks to estimate political polarization and the Perspective API to estimate toxicity, is that it is limited to relatively more active users and subreddits. We are only able to develop, in line with past works, toxicity norms and political estimations for subreddits that have at least fifty comments and more than ten political URL submission posts. As a result, our results are skewed to subreddits and users that post more often. However, we argue that these subreddits and users make up a large percentage of users' experiences on the Reddit platform and thus accurately model how users interact with each other more generally. For example, our set considered subreddits has interactions from over 59.2% of all active users (posted at least once in the 18-month time frame) on the platform and nearly all of the Reddit comments and submissions.

Confounds, Correlation, and Causation. We lastly acknowledge that while we account for many user-level and subreddit-level features, there may be other hidden confounders. For example, while we attempted to remove automated accounts from much of our analysis by removing accounts that were labeled as "bot" accounts, due to the rapid rise of AI, within Reddit as a whole there could still be automated accounts. We note that we conducted this analysis for data in 2020 and 2022 before the release of ChatGPT however. We further emphasize that while we work to account for confounders, the results we present describe the correlation between misinformation, political polarization, and toxicity; we cannot ascribe causation.

8 DISCUSSION

In this work, we examined the relationship between misinformation and politically insular and toxic environments. Using previously published lists of misinformation and authentic news domains, we find that on Reddit, the comments posted in response to misinformation submissions produce toxic comments 71.4% more often. Examining how political ideology affects the increase in toxicity in response to misinformation, we find, confirming with an Exponential Random Graph Model (ERGM), that misinformation correlates with increased toxicity between users of different political leanings. Finally, utilizing a zero-inflated negative binomial model to model engagement with misinformation versus authentic news, we observe that subreddit toxicity is a major predictor of whether misinformation submissions are commented on. This contrasts with authentic news submissions, which are often ignored within more politically polarized and toxic subreddits.

8.1 Misinformation and Authentic News

Our work shows that while misinformation has much less presence on Reddit compared to authentic news (47.8K vs 787.7K posts), misinformation plays a large role on the platform. As documented by others, often millions of comments discuss and spread false information [103]. As seen in our work, estimated levels of misinformation on particular subreddits vary widely, with some highly popular subreddits seeing upwards of 80% of the submission hyperlinks to misinformation-related sites (Figures 7 and 11). In addition to misleading users, misinformation's effect on the discourse on these subreddits can often be pernicious with articles from websites known to promote misinformation increasing inter-political strife (Section 5)

8.2 Mal-Practices: Misinformation's Correlation with Toxicity

Cinelli *et al.* [25] showed that users who post under YouTube videos promoting COVID-19 conspiracy theories often utilize toxic and vulgar language. Our paper extends this work, first showing that increased misinformation levels correlate with increased incivility on Reddit. Most importantly, we show that increased toxicity often lies in conversations between users of different political ideologies when responding to information. We also find that across much of Reddit, levels of misinformation are correlated with more insular and politically one-sided conversations, while authentic news is correlated with increased discussions between users of different political ideologies.

The community norms for particular environments appear to affect how users engage with different material. As found with our zero-inflated negative binomial model, subreddit toxicity norms are also predictive of user engagement with misinformation. Misinformation, it appears, promotes and is found within toxic environments. The more toxic/uncivil likely a given environment, the more likely at least one person is to engage with misinformation or unreliable sources. However, simultaneously, in more toxic environments, where these posts most commonly appear, these same posts are less likely to gain extensive engagement and a large number of comments. This appears to reflect misinformation submissions may often have "clickbait" titles that induce readers to initially comment, but then not often thoroughly engage with material [20, 89]. In contrast, in

less toxic environments where these posts more rarely appear, if they do gain traction (*e.g.*, at least one comment), they are more likely to gain more comments. There may be a novelty effect for individuals' engagement with these types of sources.

8.3 Political Echo-Chambers, Politics Discussions and Authentic News on Reddit

Similar to past work, we find that most toxic interactions take place among users of the same political orientation [35]. Reddit specifically creates communities for like-minded people and as a result, most interactions (including both toxic and non-toxic interactions) on the platform are between people of the same political orientation. However, most interestingly, we find that as rates of more mainstream and reliable sources used within a subreddit increase, the rate of inter-party interactions also slightly increases. We thus argue that if subreddit moderators and others want to encourage less toxic and politically diverse discussions, the usage of reliable sources across the political spectrum may help. However, we note that from our negative binomial regression results, the more polarized and ideologically distinct a subreddit becomes, the less likely that authentic news articles are to see *any* interaction from Reddit users. This suggests that while the usage of more reliable sources in more subreddits leads to more healthy conversations, these posts, when submitted in polarized subreddits are less likely to generate conversations in the first place.

Our work suggests that if communities want less toxic conversations, reliance on more accurate and reliable sources may help. Similarly, if Reddit, as a whole, desires to decrease levels of political incivility and toxicity on its platform, taking a more proactive approach to policing questionable sources could help alleviate these issues. As found by Gallacher *et al.* [44], toxic online interactions between political groups often lead to offline real-world political violence. Given that misinformation appears to be correlated with and reinforces toxic interactions between different political groups, this highlights the need to research its effects and curtail its spread.

8.4 Sub-Standards/Community Norms

We have found throughout this work that subreddits interact with authentic news and misinformation differently. For example, when polarized subreddits notice a given news post, the more polarized the subreddit, the more that users interact with the news article. Even more complexly, while more toxic subreddits are more likely to interact with misinformation, there appears to be a novelty effect, with heavily toxic subreddits commenting less on misinformation than less toxic subreddits. In contrast, more toxic subreddits, while less likely to engage at all with authentic news submissions, are more likely to heavily comment on these submissions when they do notice them. We thus find complex relationships between different types of subreddits and their interactions with different types of posts. There is no one-size-fits-all approach to understanding user engagement and toxicity on Reddit. We thus argue that a subreddit/community-based approach that takes into account the community norms of the community must be taken when trying to understand the information flows within it. Similarly, in attempting to prevent engagement with misinformation on particular subreddits, understanding their toxicity norms, their political ideology, and who is posting the article within the subreddit is key. Different communities respond differently and engage differently with these posts. We thus argue that approaches that attempt to curtail misinformation (particularly on Reddit), *must* take into account the particular nuances of that community.

9 CONCLUSION

We have seen that misinformation persists across many different types of subreddits. Its spread furthermore seems to be affected by the type of community it is posted in. Misinformation appears to be more likely to gain traction when it is posted in more toxic/uncivil environments. Furthermore, the communities with large amounts of misinformation appear to be more politically insular with

more of their interactions occurring between users of similar political orientations. As users become more dissimilar within these misinformation-filled subreddits, as found with our ERGM, they are more likely to be toxic/uncivil to one another. Comparatively, subreddits with less misinformation and more authentic news, are more likely to produce less toxic/uncivil conversations between different types of political users. Our work, one of the first to examine the relationship between misinformation, toxicity, and political ideology at scale, illustrates the need to fully understand the full effect of misinformation. Not only does misinformation mislead people but it also can magnify political differences and lead to more toxic online environments.

REFERENCES

[1] 2021. Twitter. Rules enforcement. https://transparency.twitter.com/en/reports/rules-enforcement.html-2020-jul-dec.

[2] 2022. Google Jigsaw. Perspective API. https://www.perspectiveapi.com/#/home.

[3] 2022. Metrics For Reddit - Complete List Of Subreddits - Updated Weekly. https://frontpagemetrics.com/list-all-subreddits

[4] Sara Abdali, Rutuja Gurav, Siddharth Menon, Daniel Fonseca, Negin Entezari, Neil Shah, and Evangelos E Papalexakis. 2021. Identifying Misinformation from Website Screenshots. In *International AAAI Conference on Web and Social Media (ICWSM) 2021*.

[5] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, Francesc López Seguí, et al. 2020. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of medical internet research* 22, 5 (2020), e19458.

[6] Alexa Internet, Inc. 2021. Top 1,000,000 Sites. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

[7] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[8] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[9] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3528–3539.

[10] 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.

[11] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.

[12] Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Reddit news users more likely to be male, young and digital in their news preferences | Pew Research Center. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

[13] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[14] 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.

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

[16] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2022. Finding Qs: Profiling QAnon Supporters on Parler. https://doi.org/10.48550/ARXIV.2205.08834

[17] Michael A Cacciatore, Dietram A Scheufele, and Shanto Iyengar. 2016. The end of framing as we know it… and the future of media effects. *Mass communication and society* 19, 1 (2016), 7–23.

[18] Pew Research Center. 2017. The partisan divide on political values grows even wider. *Pew Research Center* (2017).

[19] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[20] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*. 15–19.

[21] 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.

[22] Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 148–157.

[23] Yun Yu Chong and Haewoon Kwak. 2022. Understanding Toxicity Triggers on Reddit in the Context of Singapore. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1383–1387.

[24] 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).

[25] 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.

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

[27] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 192–199.

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

[29] Alina Darmstadt, Mick Prinz, and Oliver Saal. 2019. The murder of Keira: misinformation and hate speech as far-right online strategies. (2019).

[30] 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.

[31] Shiri Dori-Hacohen, Keen Sung, Jengyu Chou, and Julian Lustig-Gonzalez. 2021. Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2627–2628.

[32] 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.

[33] Maeve Duggan. 2017. Online Harassment 2017 | Pew Research Center. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[34] Régis Ebeling, Carlos Abel Córdova Sáenz, Jéferson Campos Nobre, and Karin Becker. 2022. Analysis of the influence of political polarization in the vaccination stance: the Brazilian COVID-19 scenario. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 159–170.

[35] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2022. Non-Polar Opposites: Analyzing the Relationship Between Echo Chambers and Hostile Intergroup Interactions on Reddit. *arXiv preprint arXiv:2211.14388* (2022).

[36] Facebook. 2021. Transparency center. https://transparency.fb.com/policies/community-standards/bullying-harassment/datz. Accessed: 2021-10-08.

[37] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[38] Christina Fink. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs* 71, 1.5 (2018), 43–52.

[39] Amos Fong, Jon Roozenbeek, Danielle Goldwert, Steven Rathje, and Sander van der Linden. 2021. The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations* 24, 4 (2021), 606–623.

[40] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[41] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. "A Stalker's Paradise" How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[42] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2017. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

[43] Daniel Funke. 2018. Fact-checkers have debunked this fake news site 80 times. It's still publishing on Facebook. Poynter. org.

[44] John D Gallacher, Marc W Heerdink, and Miles Hewstone. 2021. Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media+ Society* 7, 1 (2021), 2056305120984445.

[45] 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.

[46] 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.

[47] 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.

[48] Dipayan Ghosh and Ben Scott. 2018. Digital deceit: the technologies behind precision propaganda on the internet. (2018).

[49] Amit Goldenberg and James J Gross. 2020. Digital emotion contagion. *Trends in Cognitive Sciences* 24, 4 (2020), 316–328.

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

[51] 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.

[52] Anatoliy Gruzd and Philip Mai. 2020. Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society* 7, 2 (2020), 2053951720938405.

[53] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9, 3 (2018), 4.

[54] 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.

[55] Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying Toxic Speech Detectors Against Veiled Toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7732–7739.

[56] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. "A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems' Coverage of the Russo-Ukrainian War. *arXiv preprint arXiv:2210.03016* (2022).

[57] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. *arXiv preprint arXiv:2205.14484* (2022).

[58] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2022. No Calm in The Storm: Investigating QAnon Website Relationships. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 299–310.

[59] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. 2023. A Golden Age: Conspiracy Theories' Relationship with Misinformation Outlets, News Media, and the Wider Internet. *ACM Conference on Computer Supported Cooperative Work* (2023).

[60] Gordon Heltzel and Kristin Laurin. 2020. Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences* 34 (2020), 179–184.

[61] Marc J Hetherington. 2008. Turned off or turned on? How polarization affects political engagement. *Red and blue nation* 2 (2008), 1–33.

[62] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying Disinformation Websites Using Infrastructure Features. In *USENIX Workshop on Free and Open Communications on the Internet*.

[63] Philip N Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2019. The IRA, social media and political polarization in the United States, 2012-2018. (2019).

[64] 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.

[65] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.

[66] 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.

[67] David R Hunter, Mark S Handcock, Carter T Butts, Steven M Goodreau, and Martina Morris. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software* 24, 3 (2008), nihpa54860.

[68] 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.

[69] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[70] Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[71] Jonas L Juul and Johan Ugander. 2021. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences* 118, 46 (2021), e2100786118.

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

[73] Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* 71, 6 (2021), 922–946.

[74] 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.

[75] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the behaviors of toxic accounts on reddit. In *ACM Web Conference*.

[76] 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.

[77] 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).

[78] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559.

[79] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[80] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 515–522.

[81] Daniela Mahl, Jing Zeng, and Mike S Schäfer. 2021. From "Nasa Lies" to "Reptilian Eyes": Mapping Communication About 10 Conspiracy Theories, Their Communities, and Main Propagators on Twitter. *Social Media+ Society* 7, 2 (2021), 20563051211017482.

[82] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[83] Durim Morina and Michael S Bernstein. 2022. A Web-Scale Analysis of the Community Origins of Image Memes. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–25.

[84] Ashley Muddiman, Shannon C McGregor, and Natalie Jomini Stroud. 2019. (Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication* 36, 2 (2019), 214–226.

[85] 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.

[86] Marius Paraschiv, Nikos Salamanos, Costas Iordanou, Nikolaos Laoutaris, and Michael Sirivianos. 2022. A Unified Graph-Based Approach to Disinformation Detection using Contextual and Semantic Relations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 747–758.

[87] Tai-Quan Peng, Mengchen Liu, Yingcai Wu, and Shixia Liu. 2016. Follower-followee network, communication networks, and vote agreement of the US members of congress. *Communication research* 43, 7 (2016), 996–1024.

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

[89] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European conference on information retrieval*. Springer, 810–817.

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

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

[92] 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.

[93] 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.

[94] 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.

[95] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

[96] Martin Ridout, John Hinde, and Clarice GB Demétrio. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 57, 1 (2001), 219–223.

[97] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[98] Daniel Romer and Kathleen Hall Jamieson. 2020. Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social science & medicine* 263 (2020), 113356.

[99] 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.

[100] 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.

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

[102] Vibhor Sehgal, Ankit Peshin, Sadia Afroz, and Hany Farid. 2021. Mutual hyperlinking among misinformation peddlers. *arXiv preprint arXiv:2104.11694* (2021).

[103] Vinay Setty and Erlend Rekve. 2020. Truth be Told: Fake News Detection Using User Reactions on Reddit. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3325–3328.

[104] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Construction of Large-Scale Misinformation Labeled Datasets from Social Media Discourse using Label Refinement. In *Proceedings of the ACM Web Conference 2022*. 3755–3764.

[105] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. COVID-19 Vaccine Misinformation Campaigns and Social Media Narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 920–931.

[106] 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.

[107] 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*.

[108] Jennifer Stromer-Galley. 2003. Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication* 8, 3 (2003), JCMC836.

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

[110] 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.

[111] 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.

[112] 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).

[113] 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.

[114] Johannes van der Pol. 2019. Introduction to network modeling using exponential random graph models (ergm): theory and an application using R-project. *Computational Economics* 54, 3 (2019), 845–875.

[115] Chris J Vargo and Toby Hopp. 2017. Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on Twitter: a congressional district-level analysis. *Social Science Computer Review* 35, 1 (2017), 10–32.

[116] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)* 13, 2 (2019), 1–22.

[117] Elliott Waissbluth, Hany Farid, Vibhor Sehgal, Ankit Peshin, and Sadia Afroz. 2022. Domain-Level Detection and Disruption of Disinformation. *arXiv preprint arXiv:2205.03338* (2022).

[118] Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. 2021. A Multi-Platform Analysis of Political News Discussion and Sharing on Web Communities. In *IEEE Conference*

on Big Data.

[119] Brian E Weeks. 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of communication* 65, 4 (2015), 699–719.

[120] Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1121–1132.

[121] Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review* (2020).

[122] Magdalena E Wojcieszak and Diana C Mutz. 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication* 59, 1 (2009), 40–56.

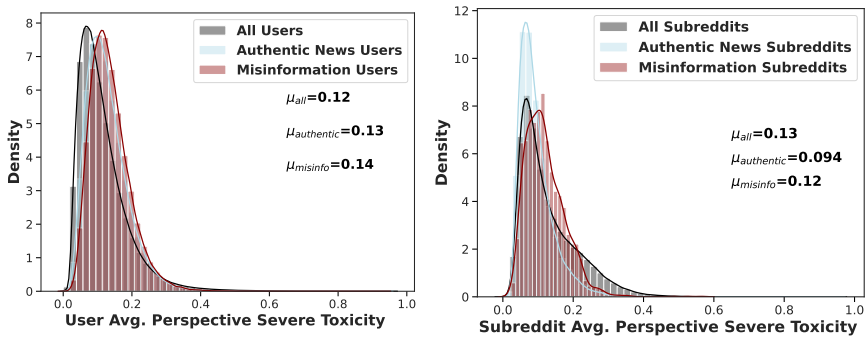
[123] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.

[124] 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.

[125] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 internet measurement conference*. 405–417.

[126] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1350–1361.

[127] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.



(a) User Toxicity

(b) Subreddit Toxicity

Fig. 14. Toxicity levels for users who comment under authentic News and misinformation URL Reddit submissions—Users who interact with misinformation submissions are slightly more toxic/uncivil than users that interact with authentic news. Both groups are slightly more toxic/uncivil than Reddit users generally. Similarly, subreddits with misinformation submissions are overall more toxic/uncivil compared with authentic news subreddits and subreddits more generally.

A ALTERNATIVE MEDIAFACT DISTRIBUTIONS

Here we present the toxicity and political ideological distribution among commenters on submissions that linked to our second set of 835 misinformation domains and 1,720 authentic news websites.

A.1 Differences in Toxicity/Incivility between Misinformation and Authentic News Submissions

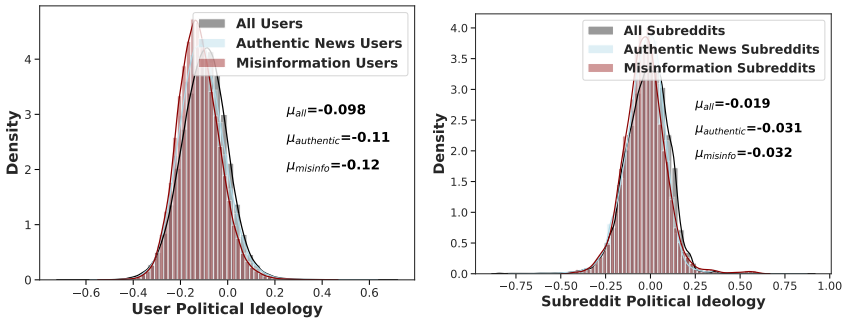
Across our second set of 9,558 misinformation and 560,673 authentic news submissions, we see a similar pattern of higher toxicity in the misinformation submission comments. 15.3% of the misinformation submissions had toxic comments with 1.25% of the comments being toxic. In contrast, 11.74% of the mainstream submissions had toxic comments with 0.64% of the comments being toxic. We thus see in this replicated experiment that Reddit misinformation conversations indeed have a higher incidence and occurrence of toxicity and incivility.

Similarly, on average 1.48% of all comments posted by the second group of misinformation commenters are toxic compared to 1.32% for the authentic news commenters (Figure 14). Looking at the subreddits where these misinformation and authentic news submissions are posted, we again see a similar trend (1.1% toxic comments vs. 0.7% toxic comments).

A.2 Differences in Political Ideology between Misinformation and Authentic News Submissions

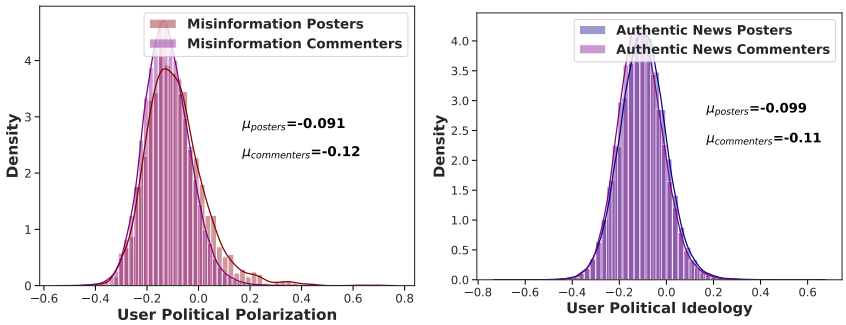
Again examining the political ideology of users commenting under misinformation Reddit submissions, we surprisingly do not see dramatic differences between them and users that comment on authentic news submissions. Similarly again looking in Figure 15 at the political orientation of the subreddits where our misinformation submissions appeared, we again see that there is not much difference in their respective political ideology distributions.

We note that despite misinformation appearing in subreddits across the political spectrum, the users that post misinformation have a rightward tilt compared to the users that comment on misinformation. As seen in Figure 16, misinformation submitters are on the whole more conservative



(a) User Political Ideology (b) Subreddit Political Ideology

Fig. 15. Political Ideology of subreddits with authentic news and misinformation Reddit submissions— There are no significant differences in political ideology between users who comment on misinformation and those that comment on authentic news. Similarly, there are no significant differences in the political orientation of subreddits where misinformation and authentic news appear.



(a) Misinformation (b) Authentic News

Fig. 16. Political Ideology of posters and commenters of misinformation— There is a noticeable rightward tilt in users who post misinformation compared to those who comment on misinformation. Unlike misinformation posts, the posters and the commenters on authentic news share similar distributions of political ideology.

than their corresponding more liberal commenters. This again is largely in contrast to authentic news commenters and posters.