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 misinformation articles are 71.4% more likely to be toxic than comments responding to authentic news articles. Identifying specific instances of commenters' 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 than in other settings. 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 Reddit 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, 45, 46, 49]. While separate and distinct phenomena, misinformation, toxic language, and political polarization are all factors that often combine with one another, stoking division and negatively affecting social media platforms [17, 25, 29, 31, 48, 80, 105, 111, 115]. While several works have attempted to understand the impact of these individual factors on social media, in this work, we explore their interaction with one another. Specifically, we seek to understand how misinformation promotes toxic and politically charged conversations across different types of online communities, asking the following three research questions:

(1) *How do online toxicity and political ideology (i.e., how conservative/Republican or liberal/Democratic) norms in a given community correlate with the presence of misinformation and unreliable news sources?*
(2) *Does political ideology help inform why users across different communities are uncivil or toxic with one another? How is this exacerbated or affected by the presence of misinformation?*
(3) *Do online toxicity and political ideology community norms affect how and when users interact with misinformation and unreliable news?*

To answer these questions, we measure the levels of toxicity, political polarization, and misinformation within submissions and comments of 18 months (January 2020 to June 2021) of data from the social media platform Reddit. Specifically, we measure the number of toxic comments within each community and by individual users on Reddit using the Google Jigsaw API [2], an out-of-the-box classifier for identifying uncivil and toxic language (e.g., insults, sexual harassment, and threats of violence [109]). Then utilizing a hyperlink-based approach as outlined by Saveski et al. [99], we approximate the political orientations of a subset of subreddits (Reddit communities) and users along the US left-right political spectrum. Finally, having approximated the toxicity

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.

and political norms within each subreddit, we utilize a curated list of misinformation websites to determine the levels at which these communities and users post misinformation. From these calculations, we tackle our three research questions:

RQ1: Toxicity and political ideology in Reddit misinformation submissions. Utilizing our list of misinformation outlets, a separate list of authentic news websites, and our pre-calculated toxicity and political norms of users and subreddits, we first determine whether there are distinct levels of user political polarization and toxicity in the comments on submissions of news articles from misinformation and authentic news sources. We find that the comments in response to 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). Commenters that respond to misinformation and authentic news come from across the political aisle. However, we observe a dichotomy in the political polarization of users that post misinformation and those that comment on misinformation: misinformation posters are more conservative than commenters. Finally, examining misinformation's correlation with overall toxicity within particular subreddits, we find that as levels of misinformation increase in a subreddit, the average toxicity of comments is also higher ($\rho = 0.352$).

RQ2: Misinformation's correlation with inter-political strife. Having identified that misinformation commenters are more likely to post toxic comments than those who respond to authentic news stories, we examine the role of political polarization in these toxic interactions. We observe that subreddits with higher amounts of *misinformation-oriented* materials are more likely to have more intra-party and insular interactions relative to subreddits with more *authentic news-oriented* materials. We further observe that commenters that post underneath misinformation submissions are more likely to be toxic to other users of different political beliefs (odds ratio 1.63). Utilizing an exponential random graph model to confirm this finding, we show that indeed, compared to other Reddit users, misinformation commenters are more likely they are to respond to users of different political views in a toxic manner. Similarly, approximating the levels of misinformation within subreddits, we find that as subreddits have more misinformation-related content posted, the more likely users of different political orientations are to be toxic to one other.

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 user toxicity and polarization levels affect 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 polarized, their users are more likely to they are to comment on misinformation 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 interactions. Our work, one of the first to examine the relationship between misinformation, toxicity, and political polarization, illustrates the need to fully understand the confluence of misinformation, polarization, and toxicity across platforms beyond Reddit. We hope that this work helps inform future research into how misinformation and unreliable sources negatively affect the health of online communities.

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 polarization factors on social media.

2.1 Terminology

Increasingly the role of social media in promoting misinformation-heavy, toxic, and highly politically polarized ecosystems has been intensely studied [24, 51, 57, 109]. Utilizing prior work, we first provide several key definitions that help to operationalize our study.

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, 52, 57, 64, 69, 77, 118]. 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, 57, 61, 85, 103, 124]. 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 [57, 61, 124].

Online Toxicity and Incivility: Given our use of the Google Jigsaw Perspective API [2], we define online toxicity and incivility as it does: "*(explicit) rudeness, disrespect or unreasonableness of a comment that is likely to make one leave the discussion.*" Within the conversations that we analyze that center around news and misinformation, we thus consider comments that meet this definition to be toxic/uncivil.

Political Ideology/Partisan Bias: We define political ideology/partisan bias as users' and communities' place on the US left/right political spectrum [96]. We note the limitation of this definition given the variety of political views within the US. However, in line with previous work [58, 98, 99], we utilize this definition that largely fits much of US-centered political discussions in order to understand how conservative-leaning and liberal-leaning users and communities interact with one another and misinformation.

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 they are moderated both by Reddit content policies, subreddit-specific rules, as well as often implicit community norms [19, 36, 68]. These norms encompass political behaviors, tolerance to misinformation, and toxic or malign behavior (*e.g.,* cursing, personal attacks) [19, 68, 93, 119]. Weld *et al.* [119] find that subreddit norms vary widely; each subreddit has its unique value hierarchy.

2.3 Political Ideology and Polarization

People, both in real life and on the Internet, tend to associate with like-minded people [10, 11, 53, 55, 65, 71, 89]. Wojcieszak *et al.* [121] 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 [107]. Social media can thus have the benefit of exposing individuals to multiple views allowing many to interact with different types of people [11, 30, 89]. 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, 59, 71]. 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 [44, 90, 108]. 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 over 12 million users, find that partisan echo chambers were driven by the algorithms of both Facebook and YouTube. Torres *et al.* [110] find the specific Twitter behavior of "follow trains" induced highly politically polarized behavior on the platform.

In a similar vein, prior works have further found that the increased political polarization engendered by social media causes increased sharing of misinformation as well as toxic online behaviors. Imhoff *et al.* [67], for example, find that political polarization is associated with beliefs in conspiracy theories. Ebling *et al.* [33] similarly find that political partisanship levels on social media are associated with medical misinformation about COVID-19. Several other authors 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 [50, 87, 111, 112].

2.4 Misinformation

In addition to driving political polarization, online activity has been found to be one of the main drivers of the spread of misinformation. As researched and reported extensively, misinformation has increasingly become a major and distinctive aspect of the conversations on social media [8, 42, 47]. Even after controlling for cascade size, Juul and Ugander find that false information spreads deeper and wider on Twitter than true information [70]. 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, 52]. As COVID-19 spread throughout the world, misinformation and conspiracy theories became a major hurdle to curbing its spread [97, 104].

To prevent the spread of misinformation, recent research has heavily focused on tracking and stemming its flow[57, 111]. Mahl *et al.* [79], 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 [51] found that a single Tweet about how COVID-19 was a hoax, spanned an entire conspiracy theory sending 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 taken used advancements in natural language processing to identify and track misinformation. Hanley *et al.* [56], for example, utilize semantic search to identify and track Russian state-media narratives on Reddit. Fong *et al.* [38] 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, 84, 120].

2.5 Toxicity

41% of Americans and 40% of those globally have reported experiencing bullying or harassment online [32, 109]. Online toxicity takes many forms including intimate partner violence, sexual harassment, doxing, cyberstalking, coordinated bullying, political incivility, and account takeovers [40, 41, 78, 109]. Toxic comments, in particular, are one of the most common forms of hate and harassment online [109]. Similar to our definition (Section 2.1), Vargo *et al.* [114] describe toxic comments as those that utilize "extremely vulgar, abusive, or hurtful language". Muddiman *et al.* define online political toxicity [82] 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 also a key aspect of social media [28, 74, 83, 109, 122]. Facebook estimates that between 0.14% and 0.15% of all views on their platform are of toxic comments [35]. This type of incivility, in addition to damaging online conversations, has been found to also damage civil institutions [15, 112] having dangerous real-world implications. Fink *et al.* [37] 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, 35]. Researchers have further taken to performing in-depth studies on users' behavior to understand abusers and victims of abuse. For instance, Founta *et al.* [39] identify a set of network and account characteristics of abusive accounts on Twitter. Hua *et al.* [63] 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 [76, 123, 125, 126].

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

Several works, close to our study, have attempted to understand how political polarization, online toxicity, and misinformation interact. Online toxicity, for instance, has been heavily associated with increased political polarization and the use of misinformation [25, 111]. Conversely, Rajadesingan et al. [92], 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 also promoted hate, toxicity, and conspiracy theories on the platform. Chen *et al.* [21], utilizing network-based analysis, find that many misleading online videos often lead to increased incivility in their comments. Separately, Rains *et al.* [91] find that high polarization is a major factor in producing 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 [73, 75, 105]. Finally, Imhoff *et al.* [67] find that political polarization is a key aspect of people's belief in false narratives.

Despite this panoply, of research, we note however, it is fairly unclear how polarization and toxicity interact in the presence of misinformation and in different political environments. In this work, we seek to fully understand these dynamics.

3 DATASETS & METHODS

Many previous works have individually investigated misinformation, toxicity, and political polarization individually thoroughly on Reddit as well as the wider Internet; we thus rely, in several places, on previous works when compiling our datasets. In this section, in addition to giving an overview of these datasets, we give a brief overview of how we calculate the political ideology/partisan bias of social media users, how we determine the toxicity levels of tweets and comments, and finally how we measure levels of misinformation within different subreddits.

3.1 Reddit Dataset

For our work, we studied 18 months of Reddit comments and submissions from January 2020 to June 2021. To aggregate this data, we rely on Pushshift [13], a third-party API that collects and publishes monthly datasets of Reddit comments and submissions. Each comment and submission includes a timestamp, the author's username, the subreddit/community where the comment was posted, and the particular conversation thread where the comment was posted. Using this data, we reconstruct the different conversation threads for each user and each 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.[1]

3.2 Misinformation and Authentic News Dataset

To begin 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,[2] OpenSources,[3] Politifact,[4] Snopes,[5] and Melissa Zimdars[6], and Hanley et al. [58]. Our final list of misinformation outlets consists of 541 websites encompassing websites such as theconservativetreehouse.com and infowars.com [58]. Many of these misinformation websites have been documented as being a part of toxic political echo chambers [106]. Separately, our list of authentic news websites consists of 565 different websites from across the political spectrum including websites like cnn.com and dailywire.com.

We note, to verify our initial findings, we rerun several of our 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.[7] 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 overlapping websites with the original list of misinformation websites, we were left with 835 websites. As our second set of authentic news websites, we utilize a set of 1885 news websites labeled as center[8], center-left[9], and center-right[10] by Media Bias/Fact Check. After removing duplicates from our original list of 565 websites, we were left with 1720 websites in our second authentic news dataset.

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

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

As in prior work [58], 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,

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

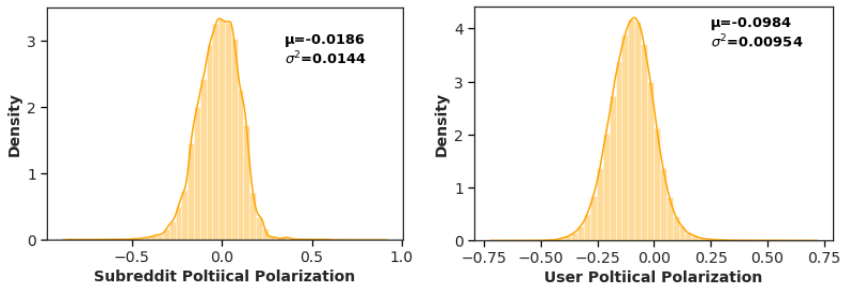


Fig. 1. Distribution of subreddit and user polarization scores — We estimate the political polarization of users and subreddits based on the political polarization of the URLs they post. We compute these estimates for users and subreddits that have posted at least 10 URLs to get robust averages for each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

we utilize Common Crawl data[11]–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 authentic news websites have hyperlink connections with which websites on the Internet; and subsequently we 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 [56, 57, 106]. Conversely, our list of *mainstream-oriented* websites includes reputable sources like nytimes.com and wsj.com [124].

3.4 Approximating the Political Polarization of Subreddits and Users

To approximate the political ideology of subreddits and users, we determine how often each user and subreddit respectively post/share conservative-leaning and liberal-leaning websites. More concretely, we utilize a dataset of website partisanship scores developed by Robertson *et al.* [96]. 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, giving each a score 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 an approximate political partisanship score of 0.0392. As found by Saveski *et al.* [98], 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.* [98] to approximate their political leanings.

[11] https://commoncrawl.org/

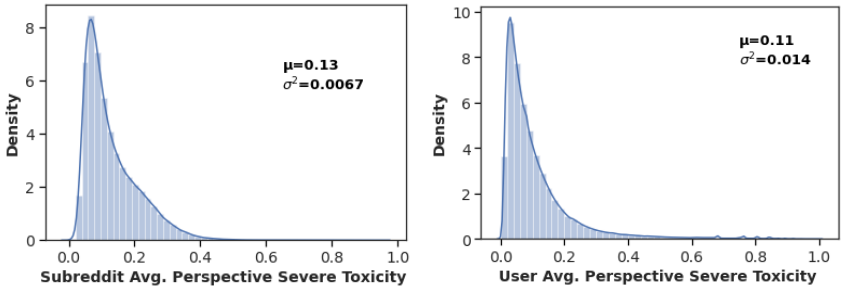


Fig. 2. Distribution of subreddit average and user average Perspective Severe Toxicity scores — We determine the toxicity norms for subreddits with at least 50 comments and users with at least 10 comments. Each user and subreddit has distinctive toxicity norms, posting toxic comments at different rates. 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.

To build a robust score for each user and subreddit, we only utilize averaged scores for 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 posted in submissions on subreddits when calculating their political leanings. We do this as these hyperlinks are implicitly approved by the given subreddit community and are more reflective of the political leanings of the full subreddit [117]. 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 (μ = 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 is only slightly liberal-leaning (μ = 0.0186). Agreeing with past work, this confirms that while subreddits are created by and for individuals across the political spectrum [93], the liberal/Democratic-leaning subreddits are the most popular and have the most users.

3.5 Identifying Toxic Reddit Comments and Approximating Users and Subreddit Toxicity Norms

to approximate the relative toxicity of Reddit users and subreddits, we utilize the Perspective API, a set of out-of-of-the-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 [74, 93, 100] and we rely on the best practices outlined in past works for our study. As in Chong *et al.*, Han et al [54] and other works, to consider a comment as toxic, we utilize a threshold of 0.8 [23, 76]. As found by Kumar *et al.* [74], 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. We note, however, that when calculating toxicity norms, we further filter down to those subreddits with at least 50 comments. For determining user toxicity norms, we first identify 31.1 million 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 [93]. 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 the distributions of these scores 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 benign 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.

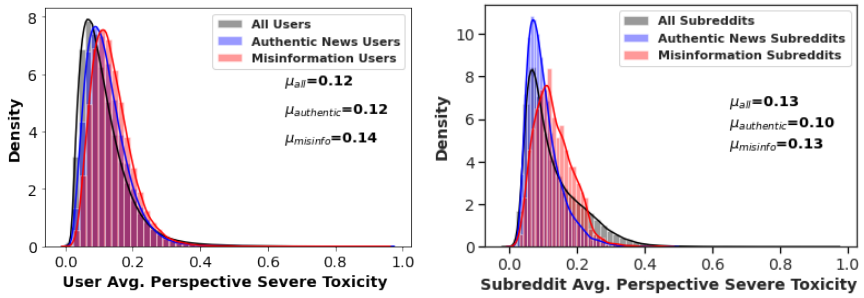
4 RQ1: REDDIT MISINFORMATION SUBMISSIONS VS AUTHENTIC NEWS SUBMISSIONS

Having outlined our methodology, in this section we examine the relative toxicity levels and political ideology of users and subreddits that interact with misinformation and authentic news.

4.1 Setup

On Reddit, as previously mentioned, users can submit news articles from websites as a submission under which users can comment. To understand the difference in levels of political polarization and toxic comments associated with posts from misinformation websites, we compare the political ideology of users and the levels of toxic comments posted under misinformation and authentic news URL submissions. Across all our collected subreddits, we gather the sets the URL submissions that utilize our set of misinformation and authentic news websites. Altogether, within our Pushshift dataset, there were 38.3K submissions utilizing our first set of 541 misinformation websites from 2.2K different subreddits and 227K submissions utilizing our set of 565 authentic news websites from 18.4K unique subreddits. The difference in the magnitude of submissions, we believe, is largely due to the greater popularity and widespread appeal of authentic mainstream news compared with alternative and more fringe websites [58]. 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.

To bolster and confirms our results, we test our findings in this section utilizing 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.



(a) User Toxicity

(b) Subreddit Toxicity

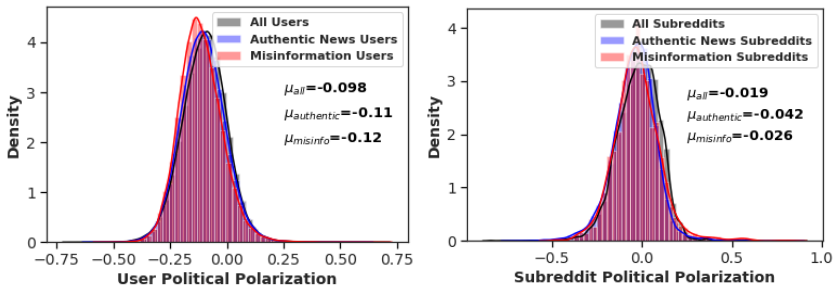
Fig. 3. Toxicity levels for users who comment under authentic News and misinformation URL Reddit submissions—Users who interact with misinformation submissions are slightly more toxic/uncivil than users who interact with authentic news. Both groups are slightly more toxic/uncivil than Reddit users generally.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

4.2 Differences in Toxicity/Incivility between Misinformation and Authentic News Submissions

Looking at the toxicity comments from submissions posting misinformation domains, we see that 14.9% of the submissions had at least one toxic comment and 1.80% of all the comments were toxic. In contrast, for our set of authentic news submissions, 13.6% of the submissions had at least one toxic comment and only 1.05% of the comments were toxic.[12] We thus see an approximate 71.4% relative uptick in the rate at which toxic comments are posted under misinformation submissions.

Higher toxicity within misinformation submissions could be caused by (1) more toxic/uncivil users participating in these conversations, or (2) higher toxicity norms in the subreddits where the misinformation was posted. As seen in Figure 3, we see that misinformation commenters are slightly more toxic than their authentic news counterparts. On average 1.54% of the comments for the users associated with misinformation submissions are toxic/uncivil compared with 1.22% for the corresponding group of authentic news users. 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 the 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 amount of toxic comments in 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

[12]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.
[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 Polarization (b) Subreddit Polarization

Fig. 4. Political polarization 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 orientation of subreddits where misinformation and authentic news appear.

comment on misinformation, are contributing to the higher rate of toxicity on misinformation submissions.

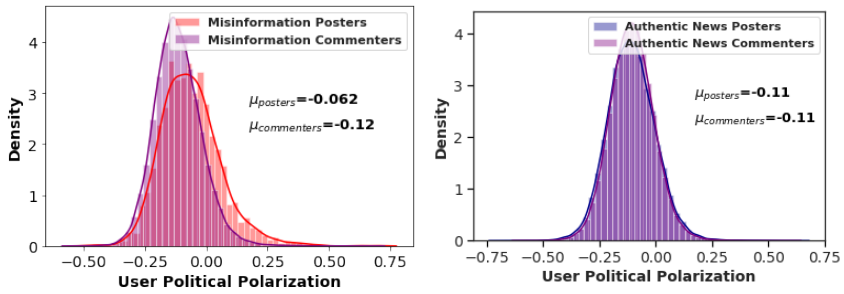
Examining the role of subreddits in promoting toxicity in Figure 3, we find that the toxicity norms of subreddits with misinformation submissions also contribute to higher levels of toxic comments. On average, the set of subreddits with misinformation submissions have higher levels of toxicity compared to subreddits with authentic news submissions. 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 that there is a heightened level of toxicity within conversations on misinformation submissions compared to authentic news submissions.

4.3 Differences in Political Polarization between Misinformation and Authentic News Submissions

Having seen the higher levels of toxicity/incivility present within misinformation Reddit submissions, we now explore the political differences between users who comment on misinformation and those that comment on authentic news.

Looking at 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. In fact, as seen in Figure 4, for our set of misinformation URLs, we see a slight leftward tilt in the average commenter. Similarly and surprisingly, looking in Figure 4 at the political orientation of the subreddits where our misinformation and authentic news submissions appeared, we again see that there is not much difference in their respective political ideology distributions. This appears to indicate that *both* misinformation and authentic news appear within subreddits and get commented on by users across the political spectrum.

We note that despite misinformation appearing in subreddits across the political spectrum, the users that post misinformation submissions have a rightward tilt compared to the users that comment on misinformation. As seen in Figure 5, we see that misinformation submitters are on the whole more conservative than their corresponding more liberal commenters. This is largely in contrast to authentic news commenters and posters. As seen in Figure 5 authentic news posters and commenters share nearly the same distribution. Altogether, we observe (especially in contrast



(a) Misinformation (b) Authentic News

Fig. 5. Distribution of the political orientation of posters and commenters of misinformation— There is a noticeable rightward tilt in users that post misinformation compared to those that comment on misinformation. Unlike misinformation posts, the posters and the commenters on authentic news share similar distributions of political orientations.

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 polarization 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 Intersection of Misinformation, News Media, Toxicity, and Political Polarization across Subreddits

Finally, having seen the political ideology and toxicity distributions among different users and in different environments, we now look if these different characteristics correlate with the amount of misinformation and authentic news in each subreddit. Namely, we now determine more broadly if levels of misinformation are correlated with increased levels of toxicity. To do this we rely on our list of *misinfo-oriented* and *mainstream-oriented* websites.

Specifically, 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.

As seen in Figure 6, across all our 46.7K considered subreddits, we observe that as subreddits become more similar to misinformation and hyperlink to more *misinfo-oriented* domains, their toxicity increases. This largely matches our observation in Section 4.2 that misinformation submissions are in general more toxic/uncivil than authentic news submissions. Misinformation levels in general thus appear to be correlated with increased toxicity. We further again surprisingly see in Figure 6 that levels of misinformation are not heavily correlated with political polarization. It does not appear that the most politically polarized environments necessarily rely upon misinformation. For example, the most left-leaning subreddits that we observed mostly supported US Senator Bernie Sanders and did not necessarily have high misinformation levels. Conversely, we do not see much of a correlation between subreddits with high mainstream similarity and political polarization and toxicity. This again reinforces our results finding that mainstream news does not have higher levels of toxicity and political polarization from Section 4.2 and Section 4.3. We thus see again from this

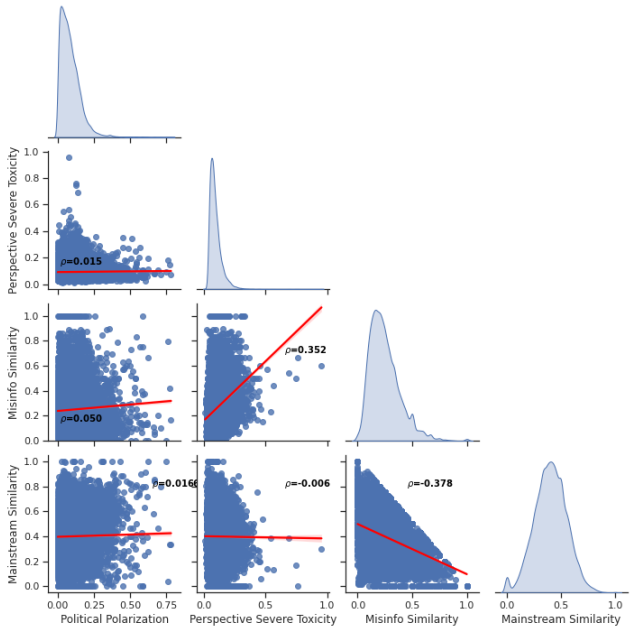


Fig. 6. Misinformation, toxicity, and political polarization interactions— As subreddits increase in misinformation levels, they become more toxic. However, there is not a large correlation between misinformation levels and the political polarization of subreddits. Similarly, we do not see any correlation between political polarization levels and mainstream similarity; nor do we see any correlation with toxicity levels. We note that given amount of subreddits considered, removing outliers with Local Outlier Detection had little effect on these results.

analysis that misinformation is indeed correlated with higher toxicity, while authentic news is largely not.

4.5 Summary

In this section, we found that misinformation on Reddit largely is correlative with and predictive of higher amounts of incivility and toxicity on the platform. 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 polarization of users that post misinformation and those that 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.

5 RQ2: MISINFORMATION AND POLARIZED TOXIC CONVERSATIONS

As seen in the previous section, comments are 71.4% more toxic in response to misinformation submissions than authentic news submissions. Furthermore, there appears to be a difference in the political orientation of users who post misinformation and those who comment on it. Given this difference and the higher toxicity levels present within misinformation submission comments, we now turn to understand if and how *political differences* are correlated with toxicity within Reddit misinformation submissions.

5.1 Setup

To fully understand how different levels of *political differences* fuel toxicity and incivility, for this section, we reconstruct the conversational dyads that exist underneath each Reddit submission using the data provided by Pushshift [13]. Comments underneath Reddit submissions are similar to

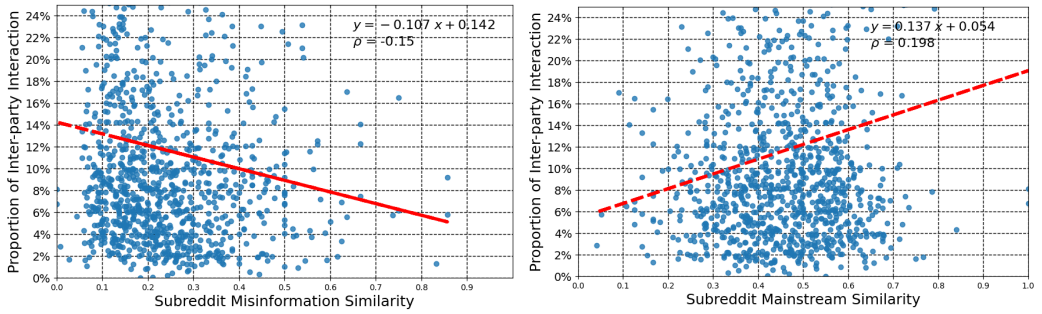


Fig. 7. Proportions of inter-party interactions in different subreddits—As subreddits hyperlink to more *misinfo-oriented* websites, as a percentage, there are fewer and fewer interactions between conservative and liberal users. In contrast, there is a slight correlation between hyperlinking to *mainstream-oriented* websites and more inter-party interactions.

conversational threads; if a user responds to a given comment, their reply will appear underneath the comment. For each submission in our dataset, we thus 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 calculations, we further label users as conservative/conservative-leaning (positive political ideology score) or liberal/liberal-leaning (negative political ideology score). Lastly, looking at each of these conversational dyads, we determine if each user's response to each other is toxic/uncivil by utilizing the Perspective API SEVERE_TOXICITY classifier with a cutoff of 0.8 (as also 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 and 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 now consider the conversational dyads under submissions that utilize our full set of 1,372 misinformation domains and 2,285 authentic news domains.

5.2 Toxic Interactions within Misinformation and Authentic News Environments

We find high amounts of homophily in interactions across our conversational dyads. Across all our conversational dyads, 81.7% of interactions are between users of the same political orientation (*i.e* liberal-liberal, conservative-conservative).[14] For conversations under authentic news[15] and misinformation submissions[16], this changes to 81.3% and 85.3% respectively. We thus see slightly more intra-party conversations within misinformation conversations than the entire Reddit population at large. Indeed using our set of *misinfo-oriented* and *mainstream-oriented* and looking at the top 1000 subreddits' conversational dyads (*i.e.* more dyads and thus more specificity in our calculated percentages), as seen in Figure 7, as website hyperlink to more *misinfo-oriented* websites, conversations on their subreddits become more insular ($\rho = -0.150$). In contrast, as subreddits

[14] Across our 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.
[15] 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.
[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.

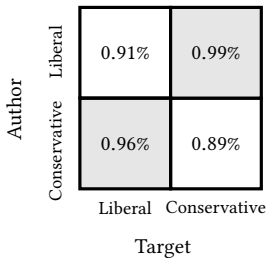


Fig. 8. Percentage of interactions that are toxic/uncivil for authors and targets of different political leanings. Across all 46K considered subreddits, there is a slight heterophily for users to reply in a toxic/uncivil manner to members with a tilt towards the opposite political party.

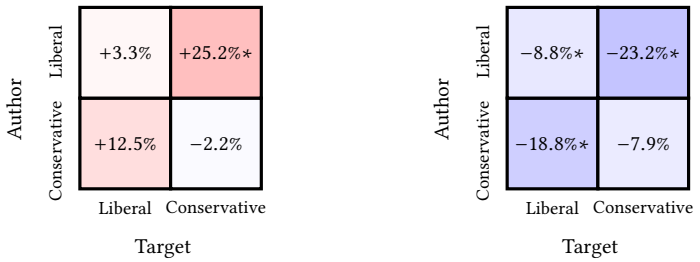
hyperlink to more *mainstream-oriented* websites, there is a slight increase in the amount of inter-party conversations (ρ = 0.198). We thus see that misinformation is correlated with heightened intra-party conversations and political homophily within the subreddits, creating more insular communities, while authentic news is associated with a slight increase in inter-party political conversations.

In contrast to increased intra-party interactions within misinformation filled-subreddits, we observe a reverse trend in terms of toxic/uncivil comments. As seen in Figure 8, across all considered conversations, we see a slight increase in users replying in a toxic/uncivil manner to users who are not of the same political leaning. We calculate an odds ratio of 1.17 for users to 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 more 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. In contrast, 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 comments (Figure 9). 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 thus find that on average, within misinformation submission comments, users are slightly more likely to respond with a toxic comment to users of different political leaning compared to all submissions on Reddit. This largely explains the higher levels of toxicity observed within misinformation submissions in Section 4.2 given the political differences we further observed between misinformation posters and misinformation comments.

Potential Confounding Factors: Account Age and Account Type. Before further examining the potential role of political ideological differences in leading to increased toxicity within misinformation submissions, we first measure the effect of some confounding factors, namely, account age and account type (Admin/Moderator vs Non-Admin/Moderator).

As suggested in prior work [94], a possible explanation for the increased toxicity within misinformation submission could be that the accounts commenting on misinformation submissions were relatively recently created (which tend to be more toxic than older accounts). Looking across all of Reddit, as seen in Figure 10, we do indeed see that newer accounts *are* more toxic than older accounts. However, upon further analysis, we see that misinformation commenting accounts tend to be older than mainstream commenter accounts *and* the relationship between age and account toxicity is less pronounced for misinformation and authentic news commenters (Figure 10). Similarly, from Table 1, we further observe that both Mod/Admin accounts (accounts that oversee



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

Fig. 9. Percentage increases of interactions that are toxic/uncivil in misinformation and authentic news submissions for conservative and liberal authors against conservative and liberal targets compared against the baseline of all interactions (Figure 8). 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.])

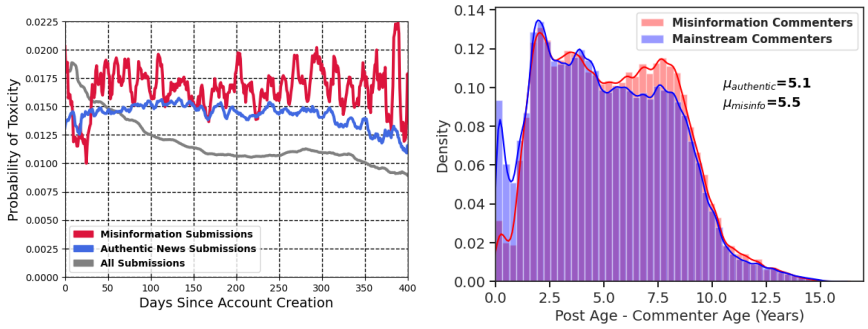
Misinformation Submissions	
Mod/Admin Toxicity	2.20% ± 0.20%
Non-Mod/Admin Toxicity	1.75% ± 0.02%
Authentic News Submissions	
Mod/Admin Toxicity	1.17% ± 0.04%
Non-Mod/Admin Toxicity	0.76% ± 0.005%
All Submissions	
Mod/Admin Toxicity	0.46% ± 0.000%
Non-Mod/Admin Toxicity	1.36% ± 0.000%

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

and regulate discussion on subreddits) have increased toxicity within misinformation submissions compared to authentic news submissions as well. We thus conclude that these possible confounding factors are not what leads to increased toxicity in misinformation submissions. However, as seen in Table 1, moderator and admin users have increased toxicity compared to non-moderator users across all news-related submissions, largely in contrast to submissions as a whole. While not responsible for the difference in toxicity between misinformation and authentic news submissions, we thus see that admin status as well as the user-creation data can have some effect on the toxicity of the users. As a result, for the rest of our analysis and or modeling, we included these factors as affecting the levels of toxicity of particular users. Having, examined these factors, we now turn to further modeling whether political ideology differences have a discriminating influence on the toxicity of misinformation commenters.

5.3 Modeling toxic interactions between users commenting under Misinformation Submissions

Having explored potential confounding factors, to further confirm the finding that users of different political stripes in misinformation-laced conversations are more likely to reply in a toxic manner



(a) Creation Date vs Toxicity

(b) Age of Commenters

Fig. 10. Newer Accounts are more likely to be toxic— Across all of Reddit, the longer that a user's account has been on Reddit the less likely they are to be toxic in any given interaction. For misinformation and authentic news submissions, this relationship, however, is less pronounced. Furthermore, on average, misinformation commenter accounts tend to be slightly older than mainstream commenter accounts. This suggests that account age is not a driving factor in users commenting in a toxic manner on misinformation submissions at a greater rate.

Misinformation Interactions	Coefficient	Mainstream/Authentic News Interactions	Coefficient
Intercept	-8.512***	Intercept	-8.711***
Absolute User Polarization	0.434	Absolute User Polarization	0.882*
User Polarization Differences	-0.893*	User Polarization Differences	-1.278***
User Toxicity	12.410***	User Toxicity	9.003***
User Moderator/Admin Status	-0.3067	User Moderator/Admin Status	-0.4660
Relative User Age (years)	0.0028*	Relative User Age (years)	-0.0049
Reciprocity	4.573 ***	Reciprocity	3.569***
Shared Subreddits	0.00063	Shared Subreddits	0.0004
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$		$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 2. 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.

to each other, we fit our network data of toxic interactions to an exponential random graph model (ERGM). An Exponential Random Graph Model (ERGM) is a form of modeling that predicts connections (*e.g.*, toxic interactions) between different nodes (users) in a given network [66]. 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.* [86], 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) [66, 113]. Utilizing this framework, we thus model the probability of toxic interactions between a given author and target 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 either 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.

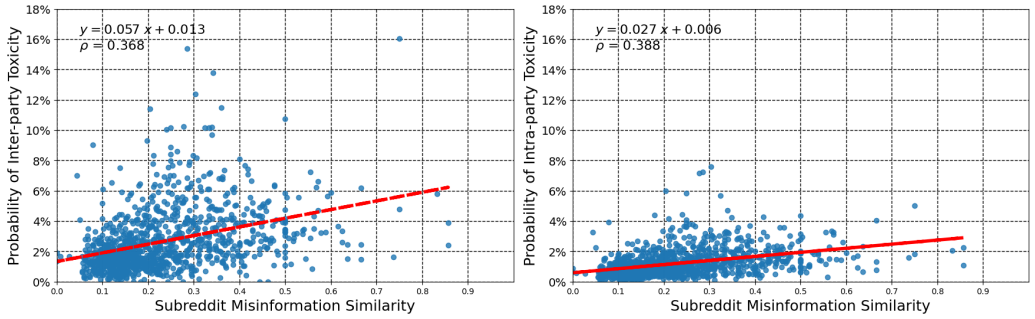


Fig. 11. Subreddit misinformation similarity vs. probability of toxic interactions between users of different and same political orientation— While for both inter-political and intra-political interactions, as misinformation similarity in a subreddit increases, the probability of a toxic interaction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

Fitting our ERGM to both misinformation submission conversational dyads and authentic news conversational dyads, we again see for both misinformation interactions and authentic news interactions that account age and account admin/moderator status, have little to no significant effect on the interactions in both cases. Indeed for misinformation interactions, we see that as accounts get *older* relative to when they post the more likely they are to engage in toxic interactions. From Table 2, for authentic news interactions, we further observe (1) that the more toxic a user, the more likely they are to engage in toxic interactions, (2) that users are more likely to respond in a toxic manner to users who engage with them in a toxic manner (reciprocity), (3) that users who are *more* politically similar to each other are more likely to be toxic to each other (*i.e.* there are more interactions amongst users who are politically similar than different), and finally (4) that more polarized and ideological a user is (on either side of the political spectrum) the more likely they are to engage in toxic behavior. These results largely concur with our previous discussions. Directly comparing misinformation interactions, we see mostly similar behavior, with toxic conversations being more likely to occur among toxic users and with users more likely to respond in a toxic manner to others who send them a toxic comment first. However, most importantly, we find here, while most toxic interactions do still 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.

5.4 Levels of Misinformation and increased rates of inter-political toxicity

Finally, having confirmed that users posting under misinformation submission of different political orientations are more likely to engage in negative interactions with each other, we determine if the overall levels of misinformation within given subreddits as a whole leads to increased inter-party toxic interactions. Namely, as misinformation levels in a subreddit as a whole increase does the probability of negative interactions between users of different political orientations increase? We

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 3. 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)

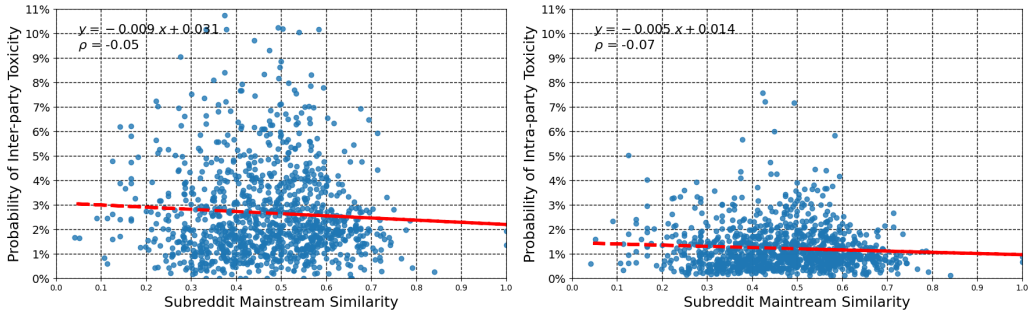


Fig. 12. Subreddit misinformation similarity vs. probability of toxic interactions between users of different and same political orientation— For both inter-political and intra-political interactions, as mainstream similarity in a subreddit increases the probability of inter- and intra-political toxicity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

thus plot the percentage of misinformation within a given subreddit against the probability of toxic interaction between members of the two political orientations.

As seen in Figure 11 looking at subreddits with at least 50 toxic conversational dyads (*i.e.* subreddits with a large number of interactions), we see that as subreddits have more *misinformation-oriented* hyperlink submissions, the percentage of conversations dyads between users of different political leanings increases. Concretely, subreddit *misinformation similarity* and the probability with which inter-political conversations are toxic correlate $\rho = 0.368$. While we similarly see that intra-political toxicity as a function of the amount of *misinformation-oriented* hyperlinks also increases with a similar correlation $\rho = 0.388$, we see 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 in Table 3 do indeed see the inter-party toxic increases at a faster rate than intra-party interactions. This reflects the fact that *misinformation oriented* submissions are on the whole more toxic but that they increase inter-party toxicity at a faster rate than politically 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. Similarly, after fitting our linear regression and performing the same moderation analysis, in Table 4 we find, as in Section 5.2, that mainstream similarity correlates with a decreased rate of interparty party toxicity (-0.00397).

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 4. 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

5.5 Summary

Throughout this section, we find that hyperlinks from misinformation outlets not only promote higher levels of toxicity in general but also appear to be correlated with increased inter-political incivility. Fitting an ERGM to our misinformation toxic dyads, we indeed find that political differences (along with reciprocity and each user's toxicity) drive toxic interactions at a higher rate. Finally, examining how different levels of misinformation promote 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 explored how misinformation on Reddit correlated with more toxic and politically insular environments on Reddit, we now determine these factors' role in user engagement with misinformation and authentic news. 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 more and engage more thoroughly with misinformation in these toxic and politically insular environments?

Various works have found that toxic and polarized environments often provoke engagement from users as they get "outraged" by the presented content [43, 72]. In this final section, we seek to determine how different communities and different community norms affect the rates at which misinformation and authentic news get interactions on Reddit.

6.1 Setup

To understand user interactions and engagement with misinformation and authentic news URL submissions, we utilize the number of comments that each submission receives. 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. We lastly note that 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). We thus consider the number of comments from "real world" users.

To model the count data of the number of comments on given submissions, we utilize a zero-inflated negative binomial regression [95]. Within our regression, each observation data point represents a single submission and the number of comments it garnered. We specifically 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

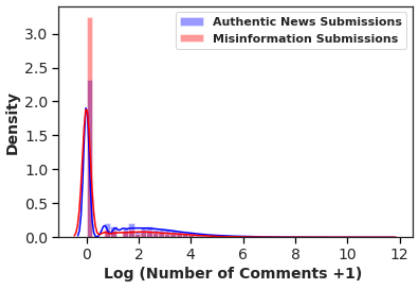


Fig. 13. Log of the number of submissions for misinformation and authentic news submissions— A large majority of submissions do not receive comments.

make the strong assumption that the mean of the data is equal to the variance [81]. Given that some submissions garner thousands of comments while others garner none, utilizing a Poisson model would be somewhat inappropriate. 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 thus 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, gives 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 perform a spot-check on the results: we ensure that the higher the average amount 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 received more comments. As seen in both Tables 5 and 6, for both misinformation and authentic news Reddit submissions, as the average number of comments in a particular subreddit increases, (1) the more likely a submission is to get comments at all and (2) the more comments it is likely to get. Having observed this behavior, we now examine the rest of the covariates within our fits.

User Admin/Moderator Status. For both misinformation and authentic news submissions, we observe that user/moderator accounts, when they post either misinformation or authentic news

submission, are less likely to garner comments. However, if these submissions eventually do gain comments, we find that if the submitter is a moderator/admin, this status is correlated with the submissions getting more comments.

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 as account age increases, that submissions are more likely to get more comments in both cases. This may indicate that accounts with more history in given subreddits

Number of Comments on Misinfo 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-Creation Date	0.021	0.045***
Absolute User Polarization	-5.635***	-2.584 ***
Absolute Subreddit Polarization	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 5. 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 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-Creation Date	0.024	0.082***
Absolute User Polarization	-3.066***	-1.212***
Absolute Subreddit Polarization	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 6. Fit of our zero-inflated negative binomial regression on the number of comments on our set of authentic news URL submissions across different subreddits.

or on Reddit as a whole may attract more engagement with their posts due to their reputation or knowledge of the platform.

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 polarized (*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 largely agrees with prior work that has shown that highly polarized users are likely to provoke and garner comments on social media platforms [62, 72].

However, despite highly polarized 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 polarized, 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 thus engage less with highly polarized users [60] compared to more politically neutral users.

Subreddit Polarization. We find that for authentic news submissions and misinformation submissions, the more politically polarized 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 generally these news posts, and in particular authentic news submissions, do not ordinarily gain any traction on highly polarized subreddits. Inundated with news, Rather, as documented by Wang *et al.* [117] subreddits like these often ignore more trustworthy sources. We thus find that the insular nature of more partisan subreddits may be inducing them to largely ignore authentic news submissions, in particular, when they are posted.

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 seem to comment and engage with submissions [72].

Subreddit Toxicity. Looking at the subreddit toxicity, 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 [106] recently published a report entitled *"Wake Up! Even The Masks Made You Sick"!*[17] As subreddits get more toxic, we thus see that their users are more likely to engage with articles such as this.

However, we further find, for authentic news submissions, that as subreddit toxicity goes up, the more comments given 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

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

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 user toxicity, we see similar behaviors for both authentic news and misinformation submissions. Most notably, as 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 outcomes [74, 100]. This result largely matches our definition of individual toxicity as comments that *are likely to make one leave the discussion* from Section 2.1.

7 LIMITATIONS

In this work, we took a large-scale approach to understand the role of misinformation in insular and toxic communities online. Given that our approach examines multiple communities simultaneously, there are some tradeoffs we make. We outline and go over some of these limitations.

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, 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 [57, 67]. 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 [99], 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 [57, 61, 101, 116], examining misinformation from a domain-based perspective enables researchers to track readily-identifiable questionable information across different platforms and thus is a reliable way of understanding the presence of misinformation.

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 largely 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 50 comments and more than 10 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 have 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.

We lastly acknowledge that while we take into account many user-level and subreddit-level features throughout this work, there can of course be other confounders that we did not actively consider.

8 DISCUSSION

In this work, we examined how misinformation correlates and is found within more politically insular and toxic environments. Using lists of misinformation and authentic news domains, we find that the comments underneath Reddit submissions using misinformation websites produced toxic comments 71.4% more often. Examining how political polarization informs the increase in toxicity within these subreddits, we find, confirming with an 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 receive comments. This is in contrast to authentic news submissions which are oftentimes ignored within more politically polarized and toxic/uncivil subreddits.

8.1 Misinformation and Authentic News

As shown in this work, despite having much less presence on Reddit compared to authentic news (47.8K vs 787.7K submissions), misinformation has a large role on the platform. As documented by others, often millions of comments discuss and spread false information [102]. As seen in our work, estimated levels of misinformation on particular subreddits vary widely, with some highly popular subreddits with upwards of 80% of the submission hyperlinks being misinformation-related (Figures 11 and 6). We thus see that while misinformation outlets *are* less popular than accurate and reliable sources, indeed many subreddits indeed contain high levels of dubious information. Furthermore, in addition to misleading users, misinformation's effect on the discourse on these subreddits can often be pernicious.

8.2 Mal-Practices: Misinformation's Correlation with Toxicity

As found by Cinelli *et al.* [25], users that post under YouTube videos promoting COVID-19 conspiracy theories, often utilize toxic and vulgar language. Our work has extended their own, illustrating not only that misinformation levels correlate with increased incivility but also that this increased toxicity often lies in conversations between users of different political ideologies. We find that across much of Reddit furthermore that levels of misinformation are correlated with more insular and politically one-sided conversations, while levels of authentic news are correlated with increased discussions between users of different political ideologies.

The community norms for particular environments seem to heavily 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 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, 88]. 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 thus 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 [34]. Reddit specifically creates communities for like-minded people and as a result, most interactions (including toxic ones) on Reddit are amongst people of the same political orientation. However, most interestingly, we find that as rates of more mainstream and reliable sources used within a subreddit increases, 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 be key. 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 get *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.

This work, thus suggests that if particular subreddits and communities want less toxic conversations, reliance on more accurate and reliable sources is one approach that they could take. 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. [43], 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 more of its effects and curtail its spread.

8.4 Sub-Standards/Community Norms

We have found throughout this work furthermore that different types of subreddits interact differently with authentic news and misinformation. For example, when polarized subreddits notice a given news post, the more polarized the subreddit, the more that it interacts 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 polarization levels, 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 polarization across such a large corpus and in multiple communities, 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] Maeve Duggan. 2017. Online Harassment 2017 | Pew Research Center. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/
[33] 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.
[34] 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).
[35] Facebook. 2021. Transparency center. https://transparency.fb.com/policies/community-standards/bullying-harassment/datz. Accessed: 2021-10-08.
[36] 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*.
[37] Christina Fink. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs* 71, 1.5 (2018), 43–52.
[38] 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.
[39] 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*.
[40] 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.
[41] 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.
[42] Daniel Funke. 2018. Fact-checkers have debunked this fake news site 80 times. It's still publishing on Facebook. Poynter. org.
[43] 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.
[44] 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.
[45] 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.

[46] 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.

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

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

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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).

[56] 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).

[57] 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.

[58] 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. *arXiv preprint arXiv:2301.10880* (2023).

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

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

[61] 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*.

[62] 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).

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.
[71] Julia Kamin. 2019. *Social Media and Information Polarization: Amplifying Echoes or Extremes?* Ph. D. Dissertation.
[72] 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.
[73] 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.
[74] 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.
[75] 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).
[76] 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.
[77] 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.
[78] 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.
[79] 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.
[80] 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.
[81] 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.
[82] 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.
[83] 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.
[84] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *AAAI Conference on Web and Social Media*.
[85] 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.
[86] 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.
[87] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.
[88] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European conference on information retrieval*. Springer, 810–817.
[89] Walter Quattrociocchi, Rosaria Conte, and Elena Lodi. 2011. Opinions manipulation: Media, power and gossip. *Advances in Complex Systems* 14, 04 (2011), 567–586.
[90] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).
[91] 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.
[92] 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.
[93] 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.

[94] 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*.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

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

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

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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*.

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

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

[109] 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.

[110] 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.

[111] 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).

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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).

[117] 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*.

[118] 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.

[119] 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.

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

[121] 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.

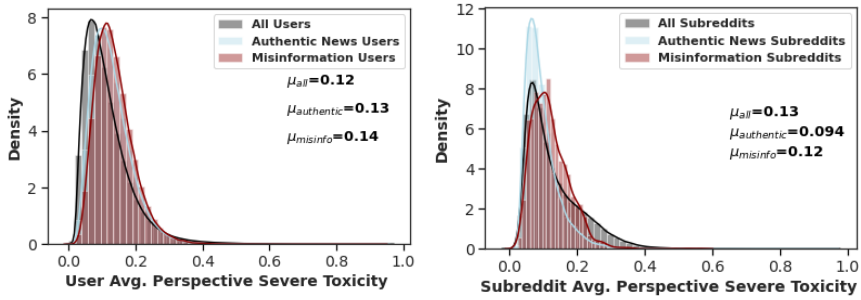
[122] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.

[123] 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.

[124] 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.

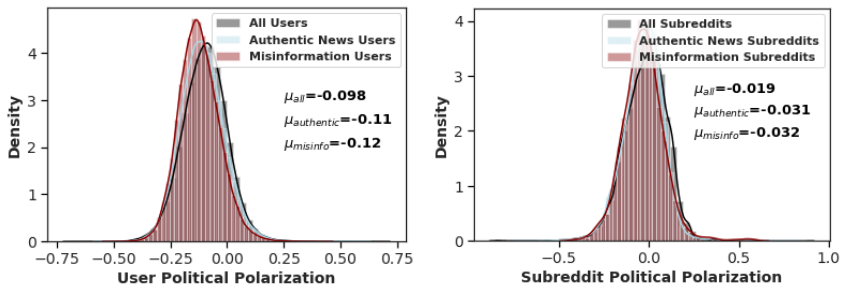
[125] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics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

[126] 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) URL Set 2 (b) URL Set 2

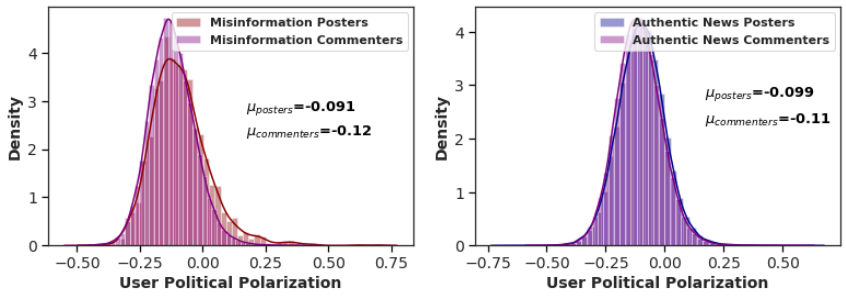
Fig. 15. Political polarization of subreddits with authentic news and misinformation Reddit submissions—There are no significant differences in political ideology between users who comment on misinformation and those that comment on authentic news. Similarly, there are no significant differences in the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

A ALTERNATIVE MEDIAFACT DISTRIBUTIONS

Here we give present the toxicity and political ideological distribution among commenters on submissions that posted our second set of 835 misinformation domains and 1720 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 again 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.



(a) URL Set 2 (b) URL Set 2

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

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 Polarization 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 again 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, we see that misinformation submitters are on the whole more conservative than their corresponding more liberal commenters. This again is largely in contrast to authentic news commenters and posters.