A Crisis of Civility? Modeling Incivility and Its Effects in Political Discourse Online

Yujia Gao[1], Wenna Qin[2], Aniruddha Murali[1], Christopher Eckart[1], Xuhui Zhou[1], Jacob Daniel Beel[1], Yi-Chia Wang[3], Diyi Yang [2]

[1] Georgia Institute of Technology, Georgia, United States
[2] Stanford University, California, United States
[3] Meta, California, United States
ygao401@gatech.edu, wennaqin@stanford.edu, amurali62@gatech.edu, chriseckart@gatech.edu, xuhui_zhou@gatech.edu, jbeel3@gatech.edu, yichia.wang@gmail.com, diyiy@cs.stanford.edu

Abstract

Growing concerns have been raised about the detrimental effects of uncivil comments on the web towards democracy. However, there is still a lack of understanding about online incivility's nuanced and complicated nature and its impact on conversation development and user behaviors. This work aims to fill that research gap by modeling incivility and its relationship to political discussions. We develop a comprehensive and fine-grained taxonomy that characterizes incivility with vulgarity, name-calling (inter-personal and third-party attacks), aspersion, and stereotypes, and then apply the framework to quantify the level of each incivility category in over 40 million comments from Reddit. Using large-scale quantitative analysis, we investigate the types of interactions and contexts in which incivility is more likely to occur, model how incivility shapes subsequent conversations, and examine user engagement patterns and behavioral changes after exposure to incivility. Our findings show that conversations that start out uncivil tend to become more uncivil in responses, and exposure to different incivility categories has differing effects on community members' engagement. We conclude with the implications of our research in assisting the design and moderation of online political communities.

Introduction

Online incivility has arisen as a crucial issue in recent years (Coe, Kenski, and Rains 2014; Cheng et al. 2017; Shmargad et al. 2022). The growth of Web 2.0, with its focus on user-generated content and end-user interactions, has progressively reshaped how people receive and participate in politics (Hmielowski, Hutchens, and Cicchirillo 2014; Rossini 2020). Consequently, political discussions in online media play a vital role in shaping democracy. While cyberspace offers grounds for opinion sharing, information exchange, and political engagement, partisan division and political polarization in recent years have become extremely far-reaching such that cross-partisan communication has been confrontational, even to a toxic extent. Hence, uncivil discourse rapidly emerges in online media, and its development will be detrimental to the construct of democracy (Santana 2014; Papacharissi 2004). One study reported that 93% of Americans identified incivility as a vast problem in society, and the Internet & social media are the second most blamed source for incivility erosion (69%), next to politicians (75%) (Shandwick 2018). To combat incivility, a growing amount of literature in recent years has studied incivility's prevalence and effect (Anderson et al. 2013; Coe, Kenski, and Rains 2014). For example, Anderson et al. (2013) showed that incivility can trigger negative reactions in those who have been directed at or exposed to it. Another line of research used empirical studies to investigate the emotional states that incivility triggers (Gervais 2015). However, there is still a lack of theoretical understanding of the more nuanced aspects of incivility and its role in online communities, i.e., what different kinds of uncivil behaviors online participants exhibit and how uncivil content influences conversations, users, and online communities in the long run.

In this work, we fill in this research gap by modeling online incivility with a fine-grained taxonomy, which deconstructs this broad concept into different categories corresponding to social, linguistic, or cultural aspects. Our ultimate goal is to develop effective design and intervention techniques to better facilitate a democratic discourse environment for political discussions. The crucial first step is gaining an understanding of how people engage in online political discussion civilly or uncivilly, into which our work deep dives from a user, conversation, and community level. We investigate the circumstances in which incivility is more likely to occur, as well as what consequences they bring to subsequent conversations. Lastly, we adopt a user-centric approach to evaluate their long-term behaviors after their exposure to incivility. In summary, we aim to address the following research questions (RQs):

1. How can nuanced aspects of incivility in online political discussions be operationalized and quantified computationally?
2. How does incivility occur in political communities, and how does its presence affect the subsequent discussion at the conversation level?
3. How does the exposure to uncivil discussions influence users' engagement with their communities?

To answer these questions, we collected around 1 million posts and 40 million comments composed by 1.6 million distinct users from Reddit. The remainder of the paper takes a deep dive into our RQs as follows: To answer RQ1, we build

out a taxonomy that accounts for different uncivil behaviors and manually annotate a curated dataset. Then, we build a set of high-performance machine learning models to predict the civility scores of each comment, which are applied to automatically label the incivility levels of the remaining comments. Next, we examine RQ2 by using statistical analyses and visualizations to understand the presence of incivility in online political communities and further investigate their effect on the health of subsequent conversations through linear regression. Finally, we look into RQ3 by performing survival analysis to study whether encounters with more incivility will affect the engagement patterns of users in the long run.

Related Work

Conversational Success and Failure

Being able to understand, quantify, and analyze the subtlety of online conversational discourse is critical to the design of regulatory mechanisms that promote a civil and democratic online environment. Beel et al. (2021) explored divisive topics on Reddit, including abortion, climate change, and gun control, to identify significant linguistic and non-linguistic features that are indicative of contentiousness, including toxicity, sentiment, and user factors. Additionally, Zhang et al. (2018) analyzed the linguistic features associated with conversational failures, such as the prompts used to start the conversation and different types of politeness strategies. On the other hand, Bao et al. (2021) aimed to define and operationalize metrics for positive social behaviors in conversation, including information sharing, gratitude, esteem boosting, and social support. Other works like Chang and Danescu-Niculescu-Mizil (2019) have studied the derailment of online Reddit conversations and forecasted their devolution into toxicity. These findings will help us understand how linguistic and social features impact subsequent discussions. Building on these works, we look closely at the concept of incivility and examine both the potential and realized impact of uncivil conversation on user engagement.

Quantifying Incivility and Similar Concepts

Extensive prior works attempted to define and predict inappropriate behaviors online, yet framed in various ways, such as explicit attacks (Wulczyn, Thain, and Dixon 2016), online abuse (Mishra, Yannakoudakis, and Shutova 2019) toxicity (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017; Zampieri et al. 2019; Pavlopoulos et al. 2022; Brassard-Gourdeau and Khoury 2019), and hate speech (Davidson et al. 2017; Schmidt and Wiegand 2017; Mosca, Wich, and Groh 2021). For example, Davidson et al. (2017) trained a multi-class model to differentiate between everyday offensive language and serious hate speech. NLP tools like the Perspective API[1], which uses machine learning models to quantify language toxicity, were launched to help researchers and developers. Despite its popularity, prior work (Hosseini et al. 2017; Jain et al. 2018) showed that an abusive comment can be slightly perturbed to deceive the system and receive a lower toxicity score even if its meaning is preserved.

To capture a broader range of uncivil behaviors, our work aligns with another line of research that defines and predicts incivility, specifically in political domains. Incivility covers a wide spectrum of intolerant behaviors, including hate speech and toxicity, but also content that breaks acceptable norms and shows disrespect. Although it seems convenient to use off-the-shelf tools like the Perspective API to predict incivility, Hede et al. (2021) pointed out that general toxicity models like Perspective API are inadequate for the analysis of incivility in news, one reason being it gives high incivility ratings for identity-related words like `black, gay, feminist, Muslim`, all of which are common topics in online political discussions. To computationally predict and quantify incivility, prior works have used logistic regression (Theocharis et al. 2020) and neural-based models (Sadeque et al. 2019; Maity et al. 2018) to predict the presence of incivility. Recently, Davidson, Sun, and Wojcieszak (2020) used a BERT-based model to classify Reddit comments as either civil or uncivil.

While our work also employs fined-tuned RoBERTa models to identify incivility, we demonstrate that online incivility is such a broad concept that can occur in different forms rooted in cultural, social, and interpersonal influences. As Kenski, Coe, and Rains (2020) found that humans perceive different kinds of incivility varies vastly, we hypothesize that the severity and effects of different uncivil behaviors on the conversations and discussion participants can also vary. As a result, a binary prediction task to classify whether incivility exists, as implemented in the previous works, is insufficient to generate a comprehensive understanding of its prevalence and effect. Thus, we make an original contribution by introducing a finer-grained incivility taxonomy to account for various dimensions of incivility to ensure a more nuanced characterization.

Impacts of Incivility

Online incivility plays a huge role in shaping conversations and online communities. Extensive experimental and large-scale quantitative work has been conducted to understand its prevalence in online media and how uncivil behaviors affect people's attitudes towards news topics (Anderson et al. 2013) or their perception of credibility for news sources (Ng and Detenber 2005). Another line of research work (Gervais 2015; Lee 2005) used randomized experiments to study the effect of exposure to incivility and found that being exposed to high levels of incivility increases feelings of anger and aversion while decreasing satisfaction. However, limited research has been conducted in a quantitative setting to explore the effects of incivility exposure on user engagement, and our work attempts to fill this gap by modeling user participation duration via causal inference methodology. On political discussions specifically, large-scale studies examined incivility in online venues where political discussions are hosted, such as online newspapers, forums, and social media (Hua, Naaman, and Ristenpart 2020). Papacharissi (2004) revealed that 36% of randomly selected posts on a political discussion board contain incivility or impolite behaviors. Su

[1]https://perspectiveapi.com

et al. (2018) classified incivility by intensity and directionality and applied their framework to large-scale Facebook political discussions, and found that extremely uncivil comments occur frequently in homogeneous discussions. Similar to our work, Xia et al. (2020) used the Perspective API to quantify the toxicity level of Reddit discussions to investigate the antecedents and consequences of toxicity. Kumar et al. (2023) also leveraged Perspective API to study the patterns of abusive accounts on Reddit. We complement the above works by using a large-scale quantitative methodology and fine-grained taxonomy to understand the role that online incivility plays in political communities on a pseudo-anonymous, content-based platform, and investigate topics like user engagement and conversation derailment that has not been looked at yet, which adds to the theoretical understanding of incivility.

Incivility Framework

In our work, we seek to detect, quantify, and analyze online incivility in political discussions. While prior work extensively investigated hate speech or offensive languages in online discussions (Warner and Hirschberg 2012; Davidson et al. 2017), we make a distinction between those with incivility, which is a broader concept that includes hate speech and offensive language but also covers social, linguistic, or cultural behaviors that violate social norms of communication and can be detrimental to a democratic discourse environment. To answer our RQ1, we create a taxonomy that operationalizes incivility and uses machine learning models to quantify each subcategory, which can be applied to large-scale political discussion data from Reddit.

As suggested by Herbst (2010), incivility is often contextual and can be subjectively perceived, which explains why there has not been a unified definition for it by scholars within this domain. One proposed definition is "*a deliberate disrespect and insult*" (Gervais 2015), while another is "*an explicit attack that insults another person's character and detracts from healthy, heated debate*" (Anderson and Huntington 2017), etc. For our taxonomy, we align with the majority of previous works that study online political discussions (Sobieraj and Berry 2011; Coe, Kenski, and Rains 2014) and prioritize broadly on behaviors that *disrespect towards individuals, groups, political communities, or topics of discussion*. Contrary to prior literature, which considers incivility as a single class, we recognize that such a generalization could miss a large potential for understanding the nuanced makeup of such a subtle concept and instead build a taxonomy that categorizes prominent aspects of incivility.

To construct the taxonomy, we initially referenced prior works' attempts to operationalize incivility into different aspects (Coe, Kenski, and Rains 2014; Sadeque et al. 2019; Papacharissi 2004) and drafted multiple aspects of incivility. Then, we conducted 4 rounds of pilot annotation tasks to select the most representative and unambiguous definitions to ensure label consistency. As a result, we deconstruct incivility into four subcategories: name-calling (further classified as interpersonal or third-party attacks), aspersion, vulgarity, and stereotypes, each of which is potentially harmful to civil, cooperative, and democratic cyberspace. Each subcategory of our final framework is able to account for the linguistic or social aspects of online incivility.

The definitions for each subcategory of incivility are shown in Figure 1. The example snippets are paraphrased from the actual dataset we collected and annotated. As the figure demonstrates, real-life conversations, especially on content-centered platforms like Reddit, are often lengthy and/or sophisticated, and uncivil content is constructed with multiple dimensions of incivility intertwined. As a result, a multifaceted taxonomy like ours can help capture the nuances within different kinds of incivility. Under our framework, each comment could potentially employ 0-4 subcategories of incivility. While the taxonomy is representative of the major and common components of incivility, we recognize that the vast scope of human language makes it difficult to conclude that we provide an exhaustive list of incivility. In the annotation process, we also include a `other` category for annotators to account for uncivil behaviors outside our taxonomy. We then consider a comment to be uncivil if it falls into any of the categories, including `other`. Future work can be done to refine definitions to cover more nuanced aspects of online incivility.

Data

Our data consist of posts and comments from Reddit about U.S. politics over the one-year time span from February 2020 to February 2021. We have chosen Reddit as our target of investigation because the pseudo-anonymity offered by the platform creates a content-directed online environment rather than a user-focused one, like Facebook or Twitter (Su et al. 2018; Theocharis et al. 2020), which facilitates opinion-sharing and chaining discussions. The discussions happen in subreddits: user-created and managed subcommunities each with its own topic, user base, and regulation policies. Reddit communities serve a diverse range of political ideologies, which gives us opportunities to study a variety of user interactions, including homogeneous and heterogeneous ones.

Dataset Statistics

To retrieve subreddits that are relevant to U.S. politics, we used phrases "*politics*", "*us politics*", and "*election*" to query subreddits. Through manual inspection, we only kept these subreddits that are relevant to U.S. politics, have public access, and have at least 20,000 subscribers. We used pushshift.io (Baumgartner et al. 2020) to retrieve 1,046,958 posts and 39,952,586 comments composed by 1,641,405 distinct users between Feb 1st, 2020 to Feb 1st, 2021. The following is a complete list of subreddits from which the data were collected:

```
r/AskTheDonald, r/AskPolitics,
r/Conservative, r/ConservativesOnly,
r/JoeBiden, r/Libertarian,
r/NeutralPolitics, r/OurPresident,
r/PoliticalDiscussion,
r/PoliticalRevolution,
r/SandersForPresident, r/TheMueller,
r/VoteBlue, r/VoteDEM, r/donaldtrump,
```

Definition	Example	% Positive	Cohen's Kappa
Aspersion is the use of disrespectful attacks or damaging statements targeted towards ideas, plans, or policies	**Dumbest argument.** You don't prove something doesn't exist, you have to prove your claim that it does exist.	9.9%	0.743
Name-calling is specific damaging or derogatory remarks towards a person, group of people, branch of the government, or political party. Name-calling is further distinguished by its target to - **Interpersonal Attacks (Towards participants in conversation)** and **Third-party Attacks (Towards Others)**	Lmao you aren't even quoting him correctly. **Your mental state is just as poor** as **Pedo Joes (Interpersonal)**	10.9%	0.810
	That's why racism exists for the majority of blacks. **Dumb white peoples** feel physically inferior to them so they knock them for being dumb even though that is stupid as hell. **(Third Party)**	17.2%	
Stereotype is the use of neutral or negative generalizations or labels, or impose discrimination upon certain groups	Well… that and **[Republican's] hatred of Muslims**, **hell** they didn't punish Oren Hazan at all for saying that they should put a china on the Muslims.	4.8%	0.940
Vulgarity is the use of vulgar language or abbreviations of such with toxic intentions towards the discussion or fellow discussants.	Can we not impeach the 63+million who voted for that **shit satan**?	17.1%	0.775

Figure 1: The definition, annotators' agreement, and examples of the subcategories of incivility in our taxonomy as well as the percent of positive samples in our annotated dataset. The colored highlighted area of the example comments demonstrates the respective incivility dimension of the same color in the definition column.

`r/hillaryclinton, r/politics, r/progressive,r/uspolitics`

Dataset Construction

To obtain ground-truth labels for different incivility categories, we sampled around 3000 comments proportional to the size of the subreddits as candidates for annotators to provide labels. To increase the representation of infrequent classes, we sample half of the labeled dataset from comments with a Perspective API score ¿ 0.8. Since incivility relies on the context it resides in, we also provide annotators with its direct parent comment and the title, i.e., topic, of the original post (OP). In addition to the incivility categories, annotators were also asked to label whether the topic of OP was divisive, defined as "*topics that evoke a mixture of positive and negative reactions*" (Hessel and Lee 2019), or "*[topics] that evoke opinionated and polarizing conversations*"(Beel et al. 2021). For example, An OP featuring divisive topic titles "*Witness corroborates claim that Lindsey Graham asked about tossing ballots in Georgia*", while an example of a non-divisive topic is "*Justice Ruth Bader Ginsburg, Champion Of Gender Equality, Dies At 87*".

We hired one research assistant who has experience with Reddit and is familiar with U.S. politics to work on this task. Moreover, for training and familiarizing the annotator with our taxonomy, we provided multiple rounds of training where the annotator was asked to label a small set of examples and to discuss the disagreements until consensus. Two authors of the work also independently annotated 10% of the dataset and inter-rater agreement among all annotators has been calculated. We used Cohen's Kappa, to measure the inter-rater agreement, shown alongside the subcategory definition in Figure 1. We recognize that differences in perception of incivility exist due to the nuance and subjectivity in perceiving conversational incivility. Thus, discussion meetings are held regularly for direct reconciliation, where the research assistant and authors of the work exchange reasoning of each annotation until unanimous decisions have been reached for all subcategories.

Detecting and Quantifying Incivility

We built machine learning models that automatically predict the incivility level of each comment in our 40 million unlabeled corpora. Since the incivility subcategories are not mutually exclusive, i.e., each comment could contain none or several types of incivility, we trained a binary classifier for each subcategory, as well as an overall incivility classification model using the aggregated incivility level of the training dataset. We experimented with four machine learning models that are widely used for text classification tasks to predict these incivility subcategories:

1. Logistic Regression (LR) with bag-of-words, Word2Vec (Mikolov et al. 2013), and TF-IDF representations respectively, implemented using `scikit-learn` package.

2. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019): a pre-trained language representation model that has outperformed other deep learning baseline methods.
3. Robustly Optimized BERT Approach (RoBERTa) (Liu et al. 2019): an enhanced BERT variation with larger amounts of training data and time, which results in state-of-the-art performances in NLP tasks. We used the Pytorch implementation of BERT and RoBERTa[2].
4. Few-shot Classification with GPT-3.5 (Brown et al. 2020): an autoregressive language model with 175 billion parameters that can achieve strong few-shot performance on many NLP datasets without any gradient updates. Specifically, we used *text-davinci-003*.

The pre-trained BERT and RoBERTa models are further fine-tuned using our annotated dataset. We use 5-fold stratified cross-validation to evaluate the performances of the first three types of models and account for potential confounding bias. For few-shot classification with the *text-davinci-003* model, we select five representative examples each from the positive and negative classes (2-way 5-shot) as the training examples and test the few-shot performance on the rest of the annotated data. [3] For each category, the prompts start with the category definition followed by 5 positive and negative examples in random order and end with an unlabeled comment query to be completed (example prompt shown in Appendix). The per-class performance of our models is listed in Table 1. As fine-tuned RoBERTa models outperformed other baselines, we applied them to the 40 million unlabeled comments. With similar model experimentation, we also utilized a fine-tuned RoBERTa model to classify the divisiveness of the OP, with model F_1 score being 0.811.

For our follow-up analyses, instead of using a binary label to represent the dimension of each incivility subcategory, we use the prediction score (range between 0 and 1) to represent the incivility subcategory score. A score closer to 1 means it is more likely to exhibit a particular uncivil category. The rationale comes from Desai and Durrett (2020), which concluded that RoBERTa is well-calibrated, meaning the probability score is a decent metric that represents the possibility of belonging to that category in real life. We evaluated the calibration of our models using *Expected Calibration Error (ECE)* (Naeini, Cooper, and Hauskrecht 2015) on the test dataset: The RoBERTa models are overall well-calibrated and have lower ECE than respective BERT experiments on most categories. The results of the calibration analyses are included in Table 6 of the Appendix.

Investigating Incivility in Conversation

This section makes use of the large-scale dataset powered by our machine learning models to investigate RQ2, which allows us to understand the development of incivility as well as its relationship with conversation development through visualization and statistical models.

[2]https://github.com/huggingface/transformers

[3]We also experimented with a 10-shot setup and found that the performances did not change dramatically.

Prevalence and Development of Incivility

Incivility is concentrated among a few highly uncivil users. From a user level, we seek to understand the onset of incivility by investigating their distribution across community members. Here, we define highly uncivil comments as those whose aggregated incivility score is greater than 85% quantile. Similar to the approach used from (Wulczyn, Thain, and Dixon 2016), we define the *incivility level* of a user as the number of uncivil comments they sent during the one-year time span of our data collection. We visualize the percentages of uncivil comments that are contributed by participants of political communities in Figure 2. Out of 589,751 highly uncivil comments, 70% of the participants with low incivility levels contributed 64.8% of uncivil comments. In the meantime, the top uncivil 6,823 (1%) users who made 100+ uncivil comments throughout the year contributed to a striking 27.8% uncivil content. In conclusion, the majority of uncivil comments were produced by a distributed group of users. However, a large portion of uncivil comments was contributed by an extremely small group of uncivil users.

Incivility is correlated with the ideology of a political community. We seek to understand whether the level of incivility is associated with the ideological leaning of political communities. Following Soliman, Hafer, and Lemmerich (2019), who used the proportion of specific leaning users as heuristics to operationalize the political leaning of subreddits, we split the subreddits into `left-leaning`, `right-leaning`, and `neutral`. G-test, a statistical technique that tests whether certain categorical data follow specific distribution, was conducted to determine whether the partisan leanings of the political communities are associated with the distribution for each aspect of incivility. The distribution for each incivility category across political slants is shown in Table 2. For all five incivility subcategories, we find significant relationships between the political leaning of the community and their prevalence. For example, the G-statistics for third-party attacks, the lowest amongst all, is 3880, $p<0.0001$. For a post-hoc two-way comparison, we have found that the level of interpersonal attacks, aspersion, stereotype, and overall incivility is the highest for conservative subreddits, whereas vulgar language and third-party attacks are most prevalent in liberal subreddits. More prominently, the level of all dimensions except for interpersonal attacks and aspersion is lowest for neutral subreddits compared to subreddits with strong partisanship leaning. This finding extends conclusions in Su et al. (2018) and indicates that more politically biased and polarized communities are more uncivil. Consistent with Soliman, Hafer, and Lemmerich (2019), we also found that right-leaning communities use more derogatory language.

Users show more incivility to others with a different political leaning. As conversations are essentially interactions between a group of two or more people, we investigate the personal dynamics of incivility between users in conversations by modeling the level of incivility with respect to the political leaning of the participants. Specifically, we compare the incivility level between two participants of the

		LR+BoW	LR+TF-IDF	LR+Word2Vec	GPT-3.5	BERT	RoBERTa
Aspersion	Acc.	0.887	**0.900***	0.900*	0.551	0.864	0.888
	Prec.	0.525	0.450	0.450	0.644	0.696	**0.738***
	Rec.	0.506	0.500	0.500	0.551	0.687	**0.728***
	F1	0.495	0.474	0.474	0.456	0.682	**0.725***
Interpersonal Attacks	Acc.	0.880	0.890	0.890	0.823	0.897	**0.902***
	Prec.	0.585	0.445	0.445	0.706	0.753	**0.759***
	Rec.	0.519	0.500	0.500	0.622	0.751	**0.768***
	F1	0.514	0.471	0.471	0.640	0.743	**0.759***
Third-party Attacks	Acc.	0.808	0.827	0.827	0.690	0.824	**0.829**
	Prec.	0.590	0.414	0.414	0.706	0.684	**0.702***
	Rec.	0.539	0.500	0.500	0.623	0.695	**0.701***
	F1	0.534	0.453	0.453	0.615	0.669	**0.679**
Stereotypes	Acc.	0.942	**0.948**	0.948	0.551	0.939	0.937
	Prec.	0.514	0.474	0.474	0.761	0.810	**0.810***
	Rec.	0.503	0.500	0.500	0.571	0.804	**0.808***
	F1	0.496	0.487	0.486	0.571	0.803	**0.806***
Vulgarity	Acc.	0.856	0.837	0.829	0.844	0.851	**0.864**
	Prec.	0.760	**0.815***	0.414	0.770	0.730	0.746
	Rec.	0.670	0.531	0.500	0.730	0.741	**0.758***
	F1	0.698	0.515	0.453	**0.747***	0.717	0.734
Incivility	Acc.	0.575	0.600	0.601	0.337	0.693	**0.745***
	Prec.	0.170	0.373	0.300	0.332	0.682	**0.710***
	Rec.	0.146	0.400	0.500	0.338	0.669	**0.694***
	F1	0.129	0.301	0.375	0.331	0.624	**0.663***

Table 1: Models' per-class evaluation metrics for predicting incivility subcategories. We highlight the model with the best performance in bold. * represents statistically significant differences at the $p<0.05$ level.

same or opposite political leaning. To estimate the political disposition of users as left-leaning or right-leaning, we adopted a heuristic similar to the ones used in Rajadesingan, Budak, and Resnick (2021), which combines the subreddits that users participate in as well as the upvote ratio of their content. We extracted 1.98 million comment pairs for which the political leaning of both parent and child comments have been identified. A two-sample t-test was used to determine if there was a difference in the incivility scores when the member responded to a user with the same political leaning v.s. when they responded to a user with the opposite political leaning. The results are shown in Table 3.

We found that both left-leaning and right-leaning participants responded more uncivilly if the other person was from the *opposite* political leaning ($p<0.0001$). In particular, right-leaning users employ significantly higher levels of all incivility subcategories except for third-party attacks towards the opposite leaning, with the most prominent, i.e., interpersonal name-calling, being 177% higher ($p<0.0001$). Similarly, left-leaning users tend to comment with higher aspersion, interpersonal attacks, stereotypes, and vulgarity levels when they respond to those with opposite political leaning. One exception is left-leaning users employ higher levels of third-party attacks with those who share similar ideology, which potentially corresponds to attacks towards "common enemies", or opposition. In conclusion, on a personal level, users tend to show more incivility towards comments from users of the opposite political leaning. Our finding aligns with prior work, which showed that users adjust their linguistic style when communicating with another with the opposite political leaning (An et al. 2019). Similarly, we found that users who interact with those with *opposing* political leanings adjust the use of different incivility subcategories, specifically by more vulgar language, interpersonal name-calling, and stereotypes.

Additionally, among comments whose parent comment had the same political disposition, we add the interaction between the political leaning of the author and the subreddit. Through two-sample t-tests, we found that users engage more uncivilly if they participate in subreddits with opposite political leaning ($p<0.0001$), specifically via higher levels of interpersonal attacks ($p<0.0001$). In contrast, users use more vulgar language ($p<0.01$) within subreddits of the same political leaning. No significant difference is observed in the level of third-party attacks, stereotypes, and aspersion.

Incivility and Conversation Dynamics

This section analyzes the local context of conversations in order to understand our RQ2, i.e., how the emergence of incivility impacts subsequent conversations. We build a series of linear regression models to investigate the relationships between levels of incivility subcategories and the later development of the conversations. The dependent variables are the breadth of the reply tree, as well as the average incivility level of the subsequent conversations. We focus on direct replies to the comments of interest because the attributes of

Category	Left	Right	Neutral
Overall Incivility	24.673%	**28.646%**	21.685%
Aspersion	6.426%	**7.873%**	7.083%
Interpersonal Attacks	5.221%	**8.929%**	6.342%
Stereotype	1.705%	**3.153%**	1.302%
Third-party Attacks	**5.793%**	5.176%	3.837%
Vulgarity	**9.519%**	9.321%	3.348%

Table 2: Level of incivility associated with the political slant: The proportion of uncivil comments within liberal, conservative, and neutral-leaning subreddits. The highest level for each category is highlighted in bold, all with p<0.001.

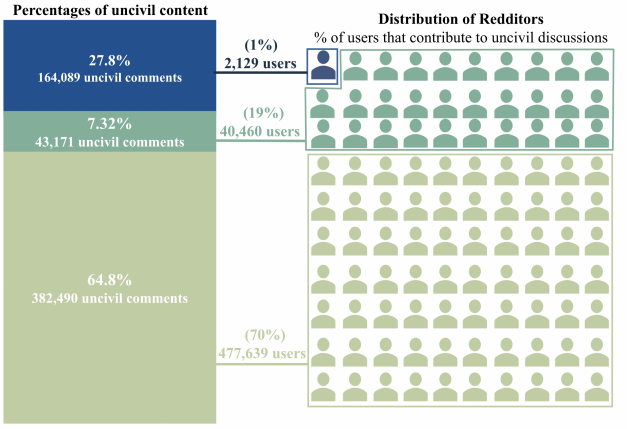

Figure 2: Distribution of Reddit users that contribute to the uncivil content of political discussions

the parent comments will mostly influence them. We filtered out comments without replies, resulting in 8.42 million comment pairs. The control variables are the divisiveness of the OP topic, sentiment, and upvote score of the top-level comments (TLC). We use VADER to operationalize the sentiment score of each comment. We inspected the correlations between the independent variables with variance inflation factors (VIF) scores and combined variables that are highly correlated with each other. The VIF score for each independent variable in our final models has a maximum of 2.93, assuring that multicollinearity was not an issue here.

Conversations that start uncivil tend to receive more responses. We use the breadth of the reply tree, i.e., the number of direct replies, as the Dependent Variable (DV) for Model 1 to measure the patterns of subsequent conversation participation. The results are in Table 4. Comments from more divisive threads receive more discussions (β_1 = 0.270). The higher the upvote score, the more direct replies it will likely receive (β_1 = 0.160). Interestingly, comments with high levels of all incivility categories except for interpersonal attacks trigger more responses to some extent, with high levels of third-party attacks with the highest coefficients. Similarly, stronger sentiment is also positively correlated with a higher number of replies (β_1 = 0.066). This implies that uncivil and highly emotional content is more likely to draw members' attention and elicit responses.

Uncivil comments are likely to elicit comments with more incivility. Model 2 in Table 4 describes how the incivility level of subsequent conversations varied with incivility aspects of comments: There is a positive correlation between the divisiveness of the OP with the incivility level of the conversations (β_1 = 0.115), which suggests that divisive topics are more likely to elicit uncivil discussions. We further verify the relationship between topic divisiveness and incivility by running t-tests against comments under divisive v.s. non-divisive posts. Results also signify comments under divisive threads have higher levels of incivility across all subcategories, with the most significant difference in `third-party attacks` (t-statistics=226.98). Comments with stronger sentiments get more uncivil responses than more emotionally neutral ones. (β_1 = 0.0195), potentially due to community members using incivility as a vent for strong emotions. In terms of the uncivil content, the positive coefficients for all incivility categories align with previous work (Gervais 2015) that incivility will beget greater incivility as conversations develop. Notably, content with higher levels of interpersonal attacks is associated with more uncivil responses than others (β_1 = 0.0593). These findings can help community moderators prioritize incivility subcategories that may be more prone to trigger more uncivil conversations.

Incivility and User Engagement

In this section, we deep dive into RQ3 and investigate the relationship between posting or being exposed to incivility and the user engagement rate of members within political communities. We use survival analysis to evaluate how members' participation is influenced by their exposure to different subcategories of incivility, the incivility level of their own comments as well as their Reddit user activities.

Survival Analysis

Survival analysis estimates the expected amount of time until an event occurs. In our case, we investigate whether exposure or the creation of incivility will affect user engagement by defining the dependent variable as the length of participation. We operationalize user engagement as the duration of participation in politics-related subreddits in units of a week. The first step to estimate the extent of exposure to incivility is to approximate the content read by users. We assume that users will read all the direct replies to their comments, as well as all the comments left in their started threads within the first week of the original posting. We acknowledge that this method might underestimate the number of comments users are exposed to because users can skim over a great amount of content on Reddit without posting. The independent variables are the level of each incivility subcategory that users are exposed to. In order to control for user characteristics that potentially confound the duration of their participation, we use community members' tenure, moderator status, activity levels, and the incivility level of their content as control variables. In our models, continuous variables were

	Incivility	Aspersion	Interpersonal attacks	Stereotype	Third-party attacks	Vulgarity
Right → Left parent	0.256	0.096	0.175	0.031	0.040	0.094
Right → Right parent	0.184	0.071	0.063	0.029	0.041	0.077
Difference	0.071***	0.024***	0.111***	0.002***	0	0.018***
Left → Right parent	0.244	0.092	0.152	0.022	0.051	0.100
Left → Left parent	0.174	0.064	0.045	0.0018	0.055	0.091
Difference	0.070***	0.028***	0.107***	0.004***	-0.004***	0.008***

Table 3: Average incivility subcategory scores for four types of interaction formed by user and parent comment pair. ***: p<0.001; **: p<0.01; *: p<0.05

Predictors	Model 1	Model 2	Predictors	Model 1	Model 2
Divisiveness of Original Post	0.270***	0.115***	Interpersonal Attack Level of PC	-0.225***	0.059***
Length of the Parent Comments (PC)	0.032***	-0.001***	Stereotype Level of PC	0.043***	0.029***
Sentiment Level of PC	0.066***	0.020***	Third-party Attack Level of PC	0.074***	0.031***
Upvote score of PC	0.160***	0.003*	Vulgarity Level of PC	0.066***	0.000
Aspersion Level of PC	0.017***	0.014***			
			Intercept	1.264	0.126
			R-squared	0.033	0.037

Table 4: Results of linear regression model predicting how the emergence of incivility impacts subsequent conversation. ***: p<0.001; **: p<0.01; *: p<0.05. The number of observations is 8,421,872. Model 1: Number of direct replies; Model 2: Average incivility level of the comment reply tree.

log-transformed and then standardized with a mean of zero and standard deviation of one, while binary variables were left in their original scale so that zero indicated an absence of the characteristic and one meant its presence. We used the statistical software package Stata to execute the analyses and assumed Weibull distribution. Note that we performed Pearson's correlation check to make sure the independent variables are not strongly correlated with each other, to prevent any multicollinearity issues.

Results The results of the survival analysis are shown in Table 5. Since the continuous variables are standardized, the Hazard Ratio (HR) represents the predicted change in the probability of leaving the community per one standard deviation increase in the predictor. An HR greater than 1 means the variable is associated with a higher-than-average likelihood of dropping out, while a hazard ratio less than 1 means a lower-than-average likelihood of dropping out.

Model 1 examines community members' likelihood to remain in Reddit politics communities based on their activities, such as tenure, posting frequencies, as well as the incivility level of the content they posted. Users who initiate more posts and comments have a 12.6% and 28.8% higher likelihood to stay in political communities, respectively. Users are also more engaged with the communities when their comments receive a higher number of upvotes (HR=0.579), which could be attributed to a sense of acceptance and belonging perceived from such social capitals within the communities (Ellison et al. 2014). Users whose comments contain higher levels of third-party attacks are more likely to stay by 2.9%. There is a possibility that participants with strong political views are motivated by their desire to express strong opinions about political figures and organizations. As pointed out by Papacharissi (2004), expression of incivility stems from strong-held opinions. The use of third-party attacks may serve as a means for community members to demonstrate their commitment to values. In contrast, participants whose comments exhibit aspersion have a 2.6% higher chance of leaving. Similarly, participants who write comments that include personal attacks towards other participants of the community are more likely to drop out by 2.3%, which can potentially be explained by moderation rules within these communities.

Model 2 adds the exposure to different subcategories of incivility. More upvotes from content read by participants increase the likelihood of staying significantly by 40.9%, indicating a greater commitment to the communities due to the exposure to these community-valued discussions. In terms of exposure to incivility, participants are 6% more likely to leave the communities when they receive comments that contain personal attacks. In other words, users who were the target or had witnessed others being attacked are more likely to leave the community, which aligns with prior work (Support and Team 2015). Similar trends exist when users are exposed to aspersion (HR=1.025). Surprisingly, exposure to content with higher levels of third-party

	Model 1	Model 2		Model 1	Model 2
Is moderator?	1.003	1.001	Length of received comments (RC)		1.058**
Tenure	0.989*	1.000	Sentiment of RC		1.0058
Number of threads initiated	0.874***	0.894***	Upvote score of RC		0.591***
Number of initialized comments (IC)	0.712***	0.782***	Incivility level of RC		0.961***
Length of IC	0.808***	0.831**	Aspersion level of RC		1.025***
Sentiment of IC	0.982***	0.985**	Interpersonal attacks level of RC		1.061***
Upvote score of IC	0.579***	0.585***	Stereotype level of RC		0.990
Incivility level of IC	0.989	0.988	Third-party Attacks level of RC		0.972**
Aspersion level of IC	1.026***	1.023	Vulgarity level of RC		0.986*
Interpersonal attacks level of IC	1.014**	1.014***			
Stereotype level of IC	0.990	0.991*			
Third-party attacks level of IC	0.971***	0.980***			
Vulgarity level of IC	1.004	1.005			

Table 5: Survival analysis predicting how long members continue to participate in political communities. ***: p<0.001; **: p<0.01; *: p<0.05. The number of observations is 113,339. The italic variables are binary variables; the rest are continuous.

attacks and vulgarity increases members' likelihood to stay at 2.8% and 1.4%, respectively. One possible explanation is that the participants were surrounded by like-minded members in *echo chambers* (Sunstein 2018). Although the received comments leveraged vulgar language and third-party attacks, they aligned with the readers' views and enabled them to feel validated and thus stay longer in the community. No significant difference is observed with exposure to stereotypes. In conclusion, exposure to uncivil discussions has varying effects on the users' long-term engagements with the political communities. Participants who send and receive uncivil content are more likely to leave the communities, with the exception of third-party attacks.

Conclusion and Discussion

This work investigates the prevalence and effect of incivility in the context of online political discussions on the social media site Reddit.com. Specifically, we create a nuanced taxonomy of representative incivility behaviors that range from cultural, social, and linguistic phenomena, including vulgarity, aspersion, name-calling (interpersonal attacks and third-party attacks), and stereotypes. With such characterizations of online incivility in political discussions, we build effective machine-learning models to predict and quantify the level of each incivility subcategory for around 40 million comments extracted from Reddit. Utilizing our fine-grained incivility framework and large-scale annotated corpus, our quantitative analyses reach the following findings that contribute to the theoretical understanding of incivility in online political discussions:

- Incivility is more prevalent in politically biased subreddits, with conservative subreddits having the highest level for all subcategories except for vulgar language and third-party attacks.
- Participants generally tend to send more uncivil comments when replying to users of the opposite political leaning, while left-leaning members tend to employ higher levels of third-party attacks when responding to members of the same leaning.
- Conversation threads that start with uncivil comments tend to gather more responses, but the conversation will also become more uncivil.
- Members are more likely to drop out from political communities when they are exposed to higher levels of aspersion and interpersonal attacks across members but tend to stay longer with exposure to more third-party attacks.

Ethical Statements

We acknowledge the potential ethical concerns in our work, especially with the dataset creation process, and take measures to mitigate the potential effects. This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution. For the dataset we collected in this work, we only leverage online media data that are publicly available at the time of data collection.

While the Reddit data are publicly available, we also recognize that the public nature of this information does not automatically imply participants' consent or comfort with their data being utilized for research purposes (Fiesler and Proferes 2018). Accordingly, we employed anonymization and data handling strategies to respect the privacy and preferences of individuals whose data might be included: To further secure personal information and user privacy, the data collection and annotations are anonymous, with user information such as username removed. Moreover, threads or comments we quoted in this work were paraphrased to prevent finding original content via web search (Bruckman 2002). Upon release of our dataset, we will only release the `id` and annotations while leaving out the comment texts themselves so that future researchers would need to re-scrape the dataset. This would at least ensure that users who might want to edit or delete their statements about personal preferences after the time of our scrape.

Broader Perspectives

Our research offers insights into methodologies of large-scale analysis of online media and resources for conducting further studies on online incivility. We introduce a comprehensive taxonomy of incivility subcategories that targets nuanced aspects of online interactions. We also annotate an incivility dataset in the political domain that offers potential for the nuanced investigation of different rhetoric of incivility. As our work suggests, the sophisticated nature of human interactions online creates a need to model the different aspects of incivility that could appear in conversations. While traditional approaches that use binary classification to detect incivility online fail to model different levels and types of uncivil behaviors, our taxonomy and methodology to quantify incivility across dimensions help fill this gap. They also open up more future research opportunities to dive deep into the concept of incivility within and beyond political contexts as well as other constructs that take root in human conversations. Moreover, We demonstrate the potential of machine learning algorithms to automatically identify different subcategories of incivility and use such techniques to create a large-scale corpus for quantitative analysis. Not only can we use the corpus to train more complicated models for better performances to better detect incivility in political subspace, but this methodology can also be applied to concepts of interest other than online incivility. The code and dataset for our work can be found at https://github.com/SALT-NLP/Incivility.

Our work also provides theoretical contributions to the understanding of incivility in online political discussions. We take a deep dive into the short and long-term effects of different subcategories of incivility on the conversations and community members' engagement. In our findings, we unraveled the nuances and complexities of the role of incivility. We discovered that exposure to some uncivil behaviors, i.e., vulgarity and third-party attacks, positively correlates with the number of replies and users' long-term engagement with the political communities. These insights suggest that while online incivility is usually perceived as harmful to conversations in online communities, certain forms of incivility are used in the political domains and even gradually keep the communities more engaged. Thus, we encourage further studies to explore incivility within the political domain to take this in mind.

Our research findings also shed light on the moderation and interventions for online media in the political domain. Aligning with previous work on other social media, our research confirms the prevalence of incivility in political discussions online, signifying the importance of effective moderation strategies. We found that a large portion of uncivil content is actually concentrated around a very small proportion of users. This observation opens up opportunities for moderators of the communities, in which potential interventions (e.g., de-platforming, shadow ban) for a small group of users might help shape a much more civil discourse environment for other users. Again, to assist with content moderation, our machine learning models prove the effectiveness of acquiring fine-grained predictions for a specific comment. Not only would the model be able to predict whether a comment is civil or not, but it also has the capacity to tell which subcategories it would be more likely to employ.

Limitations and Future Work

There are several limitations to this work. First, while our taxonomy provides a general categorization of possible acts of incivility online, the sophisticated nature of human conversation makes it impossible to study against an exhaustive characterization of such a broad concept. Examples of uncivil acts excluded in our study are lying (Coe, Kenski, and Rains 2014), misrepresented exaggeration (Sobieraj and Berry 2011), etc. Future work can build upon our and other previous work to investigate other aspects of incivility.

Another potential limitation is that most of the annotations were conducted by one research assistant, as the nuances of incivility require extensive training and interactive feedback loops to get high-quality annotations. We strive to ensure a clear understanding of our taxonomy by verifying a sample of data points and discussions to reach an agreement.

For our analysis of correlational nature, we employ a large-scale data-oriented approach to understand the rhetoric used in online incivility and its effect on shaping conversations and subsequent user engagements. While statistical methodology like survival analysis generates reasonable correlations, we recognize that incivility is influenced by other confounding characteristics, such as topics, that we do not control for and that our analysis does not provide causal effects. True causation can only be explained through random-assignment experiments.

Finally, this work focuses on the impact of incivility on political discussions within Reddit communities. Recognizing the specificity of Reddit, which may not fully mirror the broader landscape of online or offline political discourse, we acknowledge the importance of extending our investigations to other platforms and domains and look into how platform-specific features may shape conversation dynamics.

Acknowledgements

The authors would like to thank the members of the SALT group and the reviewers for their feedback. This work is funded in part by a grant from Meta and an NSF grant IIS-2144562.

References

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political Discussions in Homogeneous and Cross-Cutting Communication Spaces. *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '19*, abs/1904.05643.

Anderson, A. A.; Brossard, D.; Scheufele, D. A.; Xenos, M. A.; and Ladwig, P. 2013. The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication*, 373–387.

Anderson, A. A.; and Huntington, H. E. 2017. Social Media, Science, and Attack Discourse: How Twitter Discussions of Climate Change Use Sarcasm and Incivility. *Science Communication*, 39: 598 – 620.

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. Association for Computing Machinery.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *ArXiv*, abs/2001.08435.

Beel, J.; Xiang, T.; Soni, S.; and Yang, D. 2021. Linguistic Characterization of Divisive Topics Online: Case Studies on Contentiousness in Abortion, Climate Change, and Gun Control. *arXiv preprint arXiv:2108.13556*.

Brassard-Gourdeau, E.; and Khoury, R. 2019. Subversive Toxicity Detection using Sentiment Information. In *Proceedings of the Third Workshop on Abusive Language Online*, 1–10. Florence, Italy: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. NIPS'20. Red Hook, NY, USA: Curran Associates Inc.

Bruckman, A. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology*, 4: 217–231.

Chang, J.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1217–1230. Association for Computing Machinery. ISBN 9781450343350.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, 64: 658–679.

Davidson, S.; Sun, Q.; and Wojcieszak, M. E. 2020. Developing a New Classifier for Automated Identification of Incivility in Social Media. In *ALW*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 512–515.

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. arXiv:2003.07892.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Ellison, N. B.; Vitak, J.; Gray, R. M.; and Lampe, C. 2014. Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes. *J. Comput. Mediat. Commun.*, 19: 855–870.

Fiesler, C.; and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

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

Hede, A.; Agarwal, O.; Lu, L.; Mutz, D. C.; and Nenkova, A. 2021. From Toxicity in Online Comments to Incivility in American News: Proceed with Caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2620–2630. Online: Association for Computational Linguistics.

Herbst, S. 2010. *Rude Democracy: Civility and Incivility in American Politics*. Philadelphia, PA: Temple University Press.

Hessel, J.; and Lee, L. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. *arXiv preprint arXiv:1904.07372*.

Hmielowski, J. D.; Hutchens, M. J.; and Cicchirillo, V. 2014. Living in an age of online incivility: examining the conditional indirect effects of online discussion on political flaming. *Information, Communication & Society*, 17: 1196 – 1211.

Hosseini, H.; Kannan, S.; Zhang, B.; and Poovendran, R. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138.

Hua, Y.; Naaman, M.; and Ristenpart, T. 2020. Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates.

Jain, E.; Brown, S.; Chen, J.; Neaton, E.; Baidas, M.; Dong, Z.; Gu, H.; and Artan, N. S. 2018. Adversarial Text Generation for Google's Perspective API. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1136–1141.

Kenski, K.; Coe, K.; and Rains, S. A. 2020. Perceptions of Uncivil Discourse Online: An Examination of Types and Predictors. *Communication Research*, 47: 795 – 814.

Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2797–2807. New York, NY, USA: Association for Computing Machinery.

Lee, H. 2005. Behavioral Strategies for Dealing with Flaming in An Online Forum. *The Sociological Quarterly*, 46: 385 – 403.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Maity, S. K.; Chakraborty, A.; Goyal, P.; and Mukherjee, A. 2018. Opinion Conflicts: : An Effective Route to Detect Incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–27.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality.

Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2019. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. *ArXiv*, abs/1908.06024.

Mosca, E.; Wich, M.; and Groh, G. 2021. Understanding and Interpreting the Impact of User Context in Hate Speech Detection. In *SOCIALNLP*.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proc AAAI Conf Artif Intell*, volume 2015, 2901–2907. AAAI.

Ng, E. W. J.; and Detenber, B. H. 2005. The Impact of Synchronicity and Civility in Online Political Discussions on Perceptions and Intentions to Participate. *J. Comput. Mediat. Commun.*, 10.

Papacharissi, Z. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *SAGE Publications*, 259–283.

Pavlopoulos, J.; Laugier, L.; Xenos, A.; Sorensen, J.; and Androutsopoulos, I. 2022. From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3721–3734. Dublin, Ireland: Association for Computational Linguistics.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1125–1135. Association for Computational Linguistics.

Rajadesingan, A.; Budak, C.; and Resnick, P. 2021. Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). In *ICWSM*.

Rossini, P. G. C. 2020. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, 49: 399 – 425.

Sadeque, F.; Rains, S.; Shmargad, Y.; Kenski, K.; Coe, K.; and Bethard, S. 2019. Incivility Detection in Online Comments. *Association for Computational Linguistics*, 283–291.

Santana, A. D. 2014. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice*, 8: 18–33.

Schmidt, A.; and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *SocialNLP@EACL*.

Shandwick, W. 2018. Civility in America 2018: Civility at work and in our public squares.

Shmargad, Y.; Coe, K.; Kenski, K.; and Rains, S. A. 2022. Social Norms and the Dynamics of Online Incivility. *Soc. Sci. Comput. Rev.*, 40(3): 717–735.

Sobieraj, S.; and Berry, J. M. 2011. From Incivility to Outrage: Political Discourse in Blogs, Talk Radio, and Cable News. *Political Communication*, 28: 19 – 41.

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A Characterization of Political Communities on Reddit. *Proceedings of 30th ACM Conference on Hypertext and Social Media*.

Su, L. Y.-F.; Xenos, M. A.; Rose, K. M.; Wirz, C. D.; Scheufele, D. A.; and Brossard, D. 2018. Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, 20: 3678 – 3699.

Sunstein, C. R. 2018. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, ned - new edition edition. ISBN 9780691180908.

Support; and Team, S. 2015. Harassment Survey.

Theocharis, Y.; Barberá, P.; Fazekas, Z.; and Popa, S. A. 2020. The Dynamics of Political Incivility on Twitter. *SAGE Open*, 10.

Warner, W.; and Hirschberg, J. 2012. Detecting Hate Speech on the World Wide Web.

Wulczyn, E.; Thain, N.; and Dixon, L. 2016. Ex Machina: Personal Attacks Seen at Scale. *arXiv preprint arXiv:1610.08914*.

Xia, Y.; Zhu, H.; Lu, T.; Zhang, P.; and Gu, N. 2020. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 75–86. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Thain, N.; and Taraborelli, D. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Introduction

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes

(e) Did you describe the limitations of your work? Yes, see Limitations and Future Work

(f) Did you discuss any potential negative societal impacts of your work? Yes, see Broader Perspectives

(g) Did you discuss any potential misuse of your work? Yes, see Ethical Statements

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethical Statements

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see Broader Perspectives

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Detecting and Quantifying Incivility

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, due to the large numbers of distinct models trained, this paper presents experiment results in tables

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Detecting and Quantifying Incivility

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Detecting and Quantifying Incivility

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Analyses

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, see Broader Perspectives

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethical Statements

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethical Statements

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see Ethical Statements

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

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

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendices

GPT-3.5 Prompt Example

Prompt "Aspersion is the use of disrespectful attacks or damaging statements towards ideas, plans, or policies. The following is a list of comments and True/False labels of whether they contain aspersion.
Comment: [positive comment 1]
Has_aspersion: True
###
...
Comment: [positive comment 5]
Has_aspersion: True
###
Comment: [negative comment 1]
Has_aspersion: False
###
...
Comment: [negative comment 5]
Has_aspersion: False
###
Comment: [query comment]
Has_aspersion:
Answer: [query label]

Calibration for BERT and RoBERTa models

	BERT	RoBERTa
Overall Incivility	**0.236**	0.244
Aspersion	0.105	**0.097**
Interpersonal Attacks	0.108	**0.089**
Stereotype	0.053	**0.036**
Third-party Attacks	0.187	**0.148**
Vulgarity	0.223	**0.138**

Table 6: Expected Calibration Error (with $M = 10$ bins) on BERT and RoBERTa models per incivility category