

A Closer Look at Multidimensional Online Political Incivility

Sagi Pendzel

CS Dep., Univ. of Haifa
spendzel@campus.haifa.ac.il

Nir Lotan

IS Dep., Univ. of Haifa
nlotan02@campus.haifa.ac.il

Alon Zoizner

Dep. of Communication, Univ. of Haifa
zoizner@com.haifa.ac.il

Einat Minkov

IS Dep., Univ. of Haifa
einatm@is.haifa.ac.il

Abstract

Toxic online political discourse has become prevalent, where scholars debate about its impact on democratic processes. This work presents a large-scale study of *political incivility* on Twitter. In line with theories of political communication, we differentiate between harsh impolite style and intolerant substance. We present a dataset of 13K political tweets in the U.S. context, which we collected and labeled by this multidimensional distinction using crowd sourcing. The evaluation of state-of-the-art classifiers illustrates the challenges involved in political incivility detection, which often requires high-level semantic and social understanding. Nevertheless, performing incivility detection at scale, we are able to characterise its distribution across individual users and geopolitical regions. Our findings align with and extend existing theories of political communication. In particular, we find that roughly 80% of the uncivil tweets are authored by 20% of the users, where users who are politically engaged are more inclined to use uncivil language. We further find that political incivility exhibits network homophily, and that incivility is more prominent in highly competitive geopolitical regions. Our results apply to both uncivil style and substance. **Warning: The paper contains examples that readers might find offensive.**

1 Introduction

An active discourse between political groups and citizens is considered a fundamental condition for a healthy democracy (Gutmann and Thompson, 2009). The recent rise of social media has been argued however to intensify disrespectful and hostile online political discourse (Coe et al., 2014; Frimer et al., 2023). The term *political incivility* is commonly used in the political communication literature that studies the prominence, determinants and consequences of using harsh style and discriminatory discourse in online platforms. According

to researchers, negative consequences of this trend are that it fosters polarization between rival political groups, and may disengage some citizens from being politically involved (Muddiman et al., 2020; Skytte, 2021; Van’t Riet and Van Stekelenburg, 2022). Conversely, others argue that incivility is not inherently negative, considering certain flavors of incivility as a legitimate rhetoric in heated political discussions (Rossini, 2022). Aiming to study the consequences and contextual factors that underlie this general phenomenon, several previous research works have attempted to empirically detect, quantify and characterise political incivility on discussion groups and social media platforms (ElSherief et al., 2018; Davidson et al., 2020; Theocharis et al., 2020; Bianchi et al., 2022; Frimer et al., 2023). In this research, we take a closer look at the challenges involved in the automatic detection of political incivility online, considering it as a multidimensional concept. We then present the results of a large-scale study, where we examine incivility as detected within a very large sample of political tweets posted on the social media platform of Twitter.¹ In particular, we gauge and quantify user-level and geopolitical factors that correlate with political incivility online.

We operationalize political incivility as a two-dimensional concept in accordance with recent theories of political communication (Muddiman, 2017; Rossini, 2022). The first dimension is *personal-level incivility (impoliteness)*. This flavor of incivility pertains to a harsh tone that violates interpersonal norms, including foul language, name-calling, vulgarity, and aspersion towards other discussion partners or their ideas (e.g., “are you really so stupid that you would defund this program?”). The second dimension of *public-level incivility (intolerance)* rather pertains to exclusionary speech,

¹<https://x.com>; Our experimental data was drawn from Twitter in 2022, before this service has been re-branded as X.

IMPOLITE: “All hell has broken loose under the leadership of the senile old man. I don’t believe a damn word from this dumb son of a bitches.”; “That’s what they are protesting, you rank imbecile. People like you need a damn good kicking.”

INTOLERANT: “Hillary and the dems ARE enemies, foreign AND domestic”; “If you agree with democrats in congress, you are an anti-American commie”

NEUTRAL: “How long do Republicans believe you can keep pushing this line? You never intended to secure the border”; “There are 400,000,000 guns in the United States, you’re going to have to stop the criminals not the guns”

Table 1: Example tweets per class. These examples were presented to the annotators as part of their training.

silencing or denying the rights of a social or political group (e.g., “Democrats are openly trying to see to the destruction of America.”). That is, interpersonal incivility refers to tone, whereas intolerance is defined in terms of substance. Table 1 includes example tweets of each category. As illustrated, the impolite examples are characterized by a harsh tone, vulgar language and profanity, which may be directed at the user participating in the specific Twitter discussion (second example). However, the impolite tweets do not call for silencing an entire community or group, or denying their rights. In contrast, the example tweets of the intolerant category explicitly accuse an entire political group (in this case, Democrats) for being an enemy of the country. Regarding the neutral category, while the first example in the table criticizes Republicans, it does not call for limiting their rights or accuses their entire group of treason—thus, the tweet is not considered intolerant. Table 5 includes additional labeled examples, including a tweet that is both intolerant—as it denounces the elected U.S president along with his voters, as well as impolite—denoting its use of vulgar language. While we follow this distinction in the paper, we acknowledge that both types of incivility may be offensive. A more detailed discussion concerning the terminology of these concepts is included in Section 2.

There are several motivations for identifying political incivility at this multidimensional resolution. In general, scholars of political communication have shown that the exposure to either impolite style or intolerant content online leads to increased polarization and intergroup tensions (Muddiman et al., 2020; Skytte, 2021). Yet, recent studies argue that heated political talk should not be dismissed due to interpersonal incivility, whereas expressions of intolerance on digital platforms have a more detrimental effect on democratic processes (Pa-

pacharissi, 2004; Rossini, 2022). It is therefore desired to distinguish between the different dimensions of political incivility in studying this phenomenon. In this work, we further show that interpersonal incivility and intolerance differ in their language characteristics. While impolite speech often contains unequivocally negative lexical expressions, the interpretation of intolerance is generally a more challenging task, in that it requires contextual, political and social, understanding.

A main contribution of our work is the construction of a large dataset of 13K political tweets. We carefully retrieved and sampled these tweets using diverse strategies, aiming to capture both incivility types, while avoiding lexical and topical biases (Wiegand et al., 2019). The dataset was labeled by multidimensional incivility via crowd sourcing, having the annotation process supervised by a domain expert. Using our dataset, we adapt and evaluate a variety of state-of-the-art language models on the task of multi-label incivility detection. Our results indicate that political incivility detection is a challenging task, where we obtain best F1 scores of 0.70 and 0.59 on impoliteness and intolerance detection, respectively.

In the second part of this work, we report the results of a large scale study, in which we performed multidimensional incivility detection and examined the prevalence of incivility among the political posts by more than 200K users. We find that both types of political incivility are prevalent on social media, identifying 17.6% of the political tweets as impolite, 13.3% as intolerant and 2.5% as both, with an overall political incivility rate of 28.4%. A user-level analysis shows that a minority of the users, who are politically engaged (as measured by the proportion of their tweets that concern political topics), are more inclined to use uncivil language, generating the majority of the uncivil tweets. Our analysis further establishes that social patterns of political incivility involve network homophily. Considering the large scope of our study, we were also able to assess differences in the prevalence of incivility across geopolitical regions, specifically, states. We find that state-level incivility on social media is significantly correlated with partisan competition per state, observing higher incivility levels in ‘battleground states’, where the two camps are on par. We interpret our findings in light of existing theories of political communication, and discuss the challenges and potential of political incivility detection for future research.

2 Related work

As noted in a recent survey, the concepts of uncivil, offensive, and toxic speech often overlap, where incivility is most frequently used by social scientists (Pachinger et al., 2023). In the political communication literature, some researchers frame incivility in terms of impolite speech (Theocharis et al., 2016; Seely, 2018), whereas others define it as either impoliteness, intolerance or hate speech (Davidson et al., 2020; Theocharis et al., 2020). Accordingly, most relevant empirical studies address incivility detection as a binary classification problem, differentiating between neutral and uncivil discourse (Davidson et al., 2020; Theocharis et al., 2020; Rheault et al., 2019). Following recent theories of political communication (Rossini, 2022), we consider political incivility as a multidimensional concept, defining uncivil language as either impolite or intolerant, or both. In a closely related work, Bianchi et al. (2022) introduced a dataset of tweets annotated with fine grained labels, distinguishing between our high-level categories of rude or offensive tone (profanities, insults, outrage, or character assassination) and intolerant expressions (discrimination, hostility). Overall, they report F1 performance of roughly 0.7 on all categories. While offering valuable insights into multidimensional incivility detection, their dataset is focused on the topic of immigration, which receives limited attention in online political discourse (Barberá et al., 2019; Wojcieszak et al., 2022). Crucially, we refrained from sampling tweets based on topical keywords, while targeting political tweets by U.S. residents. Consequently, our dataset captures incivility mainly in the U.S. partisan context, which is prevalent in Twitter, across various topics (only 1.8% of the sampled tweets mention immigration). Aiming at lexical as well as topical diversity, we also minimized the use of pre-trained tools as means for sampling texts that were likely to be toxic. Possibly for these reasons, we observe substantially lower performance on intolerance detection in comparison to Bianchi et al. (F1 of ~ 0.6 vs. ~ 0.7).² Our analysis indicates that in the lack of clear lexical cues, contextual social understanding is required in order to improve on the task of intolerance detection. In this respect, our work relates to a recent line of

²We do not compare directly with (Bianchi et al., 2022), as access to their dataset was restricted at the time of this research.

works than concern the detection of implicit hate speech, where the underlying toxic intention is encoded using indirect semantics rather than by foul language (ElSherief et al., 2021; Hartvigsen et al., 2022). Finally, this work makes the contribution of applying multidimensional political incivility detection at large-scale, studying its prevalence while considering various contextual factors, including user-level characteristics and geopolitical conditions.

3 MUPID: a Multidimensional Political Incivility Dataset

3.1 Data sampling strategy

Even though political incivility is not rare, it is desired to focus the costly annotation effort on a high yield sample. We exploit multiple network-based and other cues to obtain a diverse and representative sample of the target classes, while avoiding topical and lexical biases (Wiegand et al., 2019).

As a first step, we collected tweets posted by users who follow multiple disputable political accounts, assuming that such users are more inclined to use uncivil language in political contexts (Gervais, 2014). Concretely, we referred to lists of accounts that are known to distribute fake news (Grinberg et al., 2019), news accounts that are considered politically biased to a large extent (Wojcieszak et al., 2023), and the accounts of members of the U.S. Congress who are considered as ideologically extreme (Lewis et al., 2019).³ We selected the top accounts per category, balanced over conservative and liberal orientation, based on bias scores specified by those sources.⁴ We then identified users who followed two or more biased accounts, maintaining a balance between users of conservative and liberal orientation, and retrieved the (200) latest tweets posted by them as of December 2021. This yielded 885K tweets authored by 15.8K users.

Identifying political tweets. We trained a dedicated classifier to identify tweets that discuss political topics, exploiting existing resources for this purpose. Specifically, we sampled 12.5K tweets concerning topics that are discussed frequently by either Republicans (e.g., the U.S. federal budget), Democrats (e.g., marriage equality), or both (e.g., the presidential campaign) (Barberá et al., 2015).

³<https://voteview.com/data>

⁴We selected the top ranked 20 accounts per source and orientation, except the fake news category, which includes only 9 accounts of each orientation.

Additional 3.5K political posts were extracted from the social media accounts of U.S. politicians.⁵ As counter examples, we considered random tweets by U.S. users,⁶ constructing a balanced dataset of 32K examples overall. We finetuned a ‘bert-base-uncased’ model on this dataset using its public implementation and standard training practices, minimizing the cross-entropy loss function. In applying the finetuned classifier, we set a high threshold (0.96) over its confidence scores, aiming to achieve high precision. Overall, 82K (9.3%) of our sampled tweets were predicted to be political. The manual examination of 300 random tweets by a graduate student of political communication indicated on classification precision of 0.91.

Sampling tweets for annotation. In order to focus the annotation effort on tweets that demonstrate incivility, we applied several additional sampling heuristics. Following insights by which hateful user accounts tend to be new and more active than average (Ribeiro et al., 2018), we sampled 2K tweets by accounts which were created up to two months prior to the tweet retrieval date, or posted more than one tweet daily on average since their creation date. Similar to previous works (Theocharis et al., 2020; Hede et al., 2021; Bianchi et al., 2022), we utilized the pretrained Jigsaw Perspective tool⁷ to identify toxic tweets, sampling another 2K tweets that received high scores on the categories of ‘abusive language and slurs’, ‘inflammatory comments’ and ‘attacks on the author’. Finally, we sampled 4K tweets uniformly at random. Throughout the annotation process, we tracked the yield of tweets of each class. Among the 8K selected tweets, 2.3K (28.9%) were labeled as impolite, and 0.8K (9.8%) as intolerant. Applying an active labeling paradigm (Tong and Koller, 2001), we trained a classifier of intolerance detection using the examples labeled thus far to identify additional tweets that were likely to be intolerant within our large sampled pool of political tweets. In several consequent annotation and learning batches, we selected 5.2K additional tweets for manual annotation in this fashion. The ratio of impoliteness remained similar to the original sample (22.5%), yet the ratio of intolerant tweets has tripled (29.5%). Next, we describe the annotation

⁵www.kaggle.com/datasets/crowdfLOWER/political-social-media-posts

⁶We used original tweets as opposed to retweets etc., for which the proportion of political tweets is estimated at 8% (Bestvater et al., 2022).

⁷<https://www.perspectiveapi.com/>

procedure of the sampled examples. We note that in the resulting dataset, for each example, we maintain its sampling method, where we exclude all of the examples obtained via active sampling from the test set in order to avoid evaluation bias.

3.2 Annotation procedure

The task of assessing multidimensional political incivility involves fine semantics and critical thinking. Since labeling examples by experts is costly and limited in capacity, we turned to crowd sourcing, using the platform of Amazon Mechanical Turk.⁸ In order to elicit labels of high-quality, we required the workers to be highly qualified,⁹ as well as residents of the U.S. who are presumably fluent in English and familiar with U.S. politics. Candidate workers were required to undergo dedicated training and quality testing. Table 1 includes examples which were presented to the workers of each class. These examples were accompanied by a code book containing explanations regarding the guidelines for annotating the tweets (Appendix A). In the qualification phase, the workers labeled six other tweets. Whoever labeled a majority of the tweets correctly got qualified to work on our task, as well as received detailed feedback on their mistakes. During annotation, we included control questions (2 out of 15 tweets in each micro-task) which we expected the workers to do well on. We rejected the annotations by workers who failed to label the control tweets, and banned them from further working on our task. Finally, we paid the workers an hourly fee of 17.5 USD, which exceeds the U.S. minimum wage standards, as fair pay positively affects annotation quality (Ye et al., 2017). Overall, our final cohort included 125 workers who annotated up to 2,000 tweets per week over a period of 3 months.

Given each tweet, several independent workers were asked to assess whether it was impolite, intolerant, neither, or both. Each tweet was labeled by 3-5 annotators, where we discarded examples for which a label could not be determined based on majority voting.¹⁰ While we take a prescriptive approach, we acknowledge that human judgement on this task may be subjective, being affected by one’s cultural background, beliefs, and political stance (Rottger et al., 2022). An assessment of inter-annotator agreement gives an indication for

⁸www.mturk.com/

⁹Candidate workers have completed at least 100 micro-tasks on AMT with approval rate above 98%.

¹⁰This condition is strict, as there were 4 labeling options.

Dataset	Size	Uncivil	Impol./Intol./Both
MUPID	13.1K	42.3%	24.6 / 15.1 / 2.6%
Davidson et al.	5.0K	10.3%	-
Rheault et al.	10.0K	12.4%	-
Theocharis et al.	4.0K	26.0%	-

Table 2: Dataset statistics: MUPID vs. other datasets.

the semantic complexity and subjectivity of the target concepts. Comparing the labels assigned to every tweet by random worker pairs resulted in Fleiss’ kappa scores of 0.63 and 0.54 on the categories of impoliteness and intolerance, indicating on ‘substantial’ and ‘moderate’ agreement, respectively. This suggests that intolerance may be more subjective and subtle compared to impoliteness. We further compared the majority labels against the judgement of a scholar of political communication, assigned to 300 random labeled tweets. Fleiss’ kappa scores in this case indicated on ‘substantial’ agreement, measuring 0.57 and 0.61 on impoliteness and intolerance, respectively. For a subset of this sample, for which the workers tended to agree on (majority of 70% or more), the agreement scores between the crowd sourced labels and the expert were substantially higher on the impoliteness compared to the intolerance category, measuring 0.79 vs. 0.69, respectively. Again, this suggests that the concept of political intolerance is more semantically subtle.

3.3 Dataset statistics

The resulting dataset includes 13.1K labeled tweets. As detailed in Table 2, the dataset includes a substantial number of tweets labeled as impolite (3.6K), and intolerant (2.3K), where a large proportion of the examples in the dataset (42.3%) correspond to political incivility (with 2.6% of the examples labeled as both intolerant and impolite). As noted in the table, other available datasets of political incivility use binary annotations, and include a lower proportion of examples of incivility.

4 Multidimensional incivility detection

Next, we evaluate the extent to which neural models can detect political incivility as perceived by humans. We perform multi-label classification, detecting impoliteness and intolerance as orthogonal dimensions, as well as experiment with binary prediction of political incivility.

4.1 Experimental setup

We finetuned several popular transformer-based pre-trained language models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) using our dataset. We report our results using the base configurations of these models, as the larger architectures yielded minor performance gains. In addition, we experiment with task-specialized variants of BERT: HateBERT, a model that has been re-trained using a large-scale corpus of offensive, abusive, and hateful Reddit comments (Caselli et al., 2021); and HateXplain, a model that has been finetuned to classify hateful and offensive Twitter and Gab posts (Mathew et al., 2021). All models were applied using their public implementation.¹¹ In finetuning, we split our dataset into fixed stratified train (70%), validation (10%) and test (20%) sets, optimizing the parameters of each model on the validation examples. Considering the class imbalance, we found it beneficial to employ a class-weighted cross-entropy loss function (Henning et al., 2023).

4.2 Classification results

Table 3 reports our test results in terms of ROC AUC, precision, recall and F1 with respect to each class. The table includes also the results of binary classification, considering incivility as a unified concept. As shown, binary classification yields best F1 performance of 0.75. In comparison, the best F1 results obtained for impoliteness and intolerance prediction are 0.70 and 0.59, respectively.

As baseline, we report the performance of the pre-trained Jigsaw Perspective tool, scoring the test examples by their toxicity. The Perspective model has been trained to predict toxicity as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”. Following related works, we marked as toxic the examples that received a toxicity score of 0.5 or more by the model (Gehman et al., 2020). As detailed in Table 3, this method yields high precision (0.78) yet low recall (0.43) in identifying impolite speech. Possibly, the low recall indicates on a domain adaptation issue. Toxicity is a poor predictor of intolerance however, yielding very low precision and recall scores of 0.20 and 0.18 on this category, respectively. This indicates that the intolerant examples in our dataset are not typically conveyed using general toxic language.

¹¹<https://huggingface.co/>

Classifier	Inter-personal (impolite style)				Public-level (intolerant substance)				Any incivility (binary)				
	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	MacF1
Perspective	0.841	0.781	0.432	0.556	0.674	0.200	0.180	0.189	0.850	0.897	0.329	0.481	0.636
BERT	0.857	0.635	0.713	0.671	0.848	0.530	0.644	0.581	0.849	0.752	0.692	0.721	0.766
RoBERTa	0.874	0.642	0.744	0.689	0.859	0.501	0.728	0.593	0.864	0.765	0.707	0.735	0.777
DeBERTa	0.861	0.687	0.707	0.697	0.845	0.558	0.626	0.590	0.865	0.754	0.739	0.746	0.782
HateBert	0.865	0.701	0.661	0.680	0.835	0.515	0.639	0.571	0.857	0.755	0.719	0.737	0.777
HateXplain	0.820	0.567	0.688	0.622	0.756	0.374	0.537	0.441	0.811	0.773	0.532	0.630	0.713
GPT-3.5	0.827	0.421	0.913	0.576	0.765	0.379	0.519	0.438	0.838	0.652	0.835	0.732	0.742
GPT-4	-	0.666	0.659	0.663	-	0.562	0.416	0.478	-	0.807	0.638	0.712	0.769

Table 3: Multi-label and binary prediction results.

Considering that Generative Pre-trained Transformer (GPT) models have been applied to related tasks such as hate speech detection (Wullach et al., 2021a; Del Arco et al., 2023), we further attempted few-shot incivility prediction using GPT-3.5 and GPT-4.¹² In this case, for each target category, we prompted the model with a definition of the task and category, and with (3) labeled examples that were also presented to the human workers (see Appendix A). As shown in Table 3, this approach fell short of the finetuned models. (Unlike GPT-3.5, GPT-4 no longer provides token probability information in its API. For this reason, we do not report AUC figures for GPT-4.) It is possible that further improvements in the performance of these models can be achieved via prompt engineering, additional examples or finetuning (Gül et al., 2024), however this is out of the scope of our work. Nevertheless, we observe similar trends using the GPT and the other models, showing a substantial gap in performance in favor of the impoliteness category. Concretely, we observe that GPT-4 yields F1 of 0.66 vs. 0.48 on the tasks of impoliteness and intolerance detection, respectively. The finetuned DeBERTa and RoBERTa achieve the best overall performance. Taking into account both performance and cost considerations, RoBERTa is our classifier of choice. This model yields F1 results of 0.69 and 0.59 on the impolite and intolerant classes, respectively.

Impoliteness vs. intolerance. We applied Shapley analysis (Lundberg and Lee, 2017)¹³ to our training set to identify unigrams that are predictive of impoliteness or intolerance. Table 4 lists words that characterise each class. As expected, impolite style is characterised by derogatory words. Most of the listed words carry negative meaning in an unequivocal way, being offensive in any context, e.g., ‘stupid’. In contrast, the intolerant tweets concern political affiliations, e.g., ‘republicans’, ‘right’, or

Impolite: fuck, help, stupid, damn, obnoxious, fed, joke, ass, goddamn, shit, coward, crap, unreal, love, neoliberal, king, mentality, anarchist, fuel, publishing, bad, wow, back, bastard, communists, forgive, idiot, dumb, change, worst, terrible, broke, asshole, humiliating

Intolerant: republican(s), democrat(s), leftists, GOP, democratic, catholics, speech, liberal, dem(s), socialist(s), conservatives, liberals, progressive(s), left, communist(s), party, right, racist, fascists, terrorists, nationalist(s), constituents, marxist, whites, radical, destroyed, americans

Table 4: Salient unigrams associated with impolite and intolerant speech in our dataset (Shapley analysis).

‘liberals’. Unlike slur words, negative sentiment that such terms may carry is context dependent. In accordance, we found that impolite tweets were less susceptible to get misclassified as neutral compared with intolerant tweets (26.7% vs. 44.0%). Thus, semantic and contextual understanding is needed to detect intolerance more precisely.

Error analysis. Table 5 includes examples of misclassified tweets, showing the labels assigned to them by the human workers versus the predicted labels. We indeed observe cases in which the model missed the presence of intolerance due to implied language (examples (c) and (d)), e.g., “you Republicans don’t even know how to keep the electricity on!”. Likewise, the model was sometimes misled by lexical cues, demonstrating the gap between lexical-level and semantic understanding (Zagoury et al., 2021); for instance, example (b) was misclassified as impolite, possibly because of the idiom ‘sick of’. In some other cases, we found seemingly faulty predictions to be sensible, e.g., “impeach Biden and his administration! Or charge them with treason” was justifiably classified as intolerant. Again, this demonstrates the semantic and contextual challenges involved in identifying political intolerance.

Cross-dataset evaluation. We assess learning generalization using MUPID against other relevant

¹²GPT-3.5-turbo-instruct and GPT-4-turbo, see <https://platform.openai.com/docs/models>

¹³<https://github.com/slundberg/shap>

Tweet	Label	Prediction
(a) We need to impeach Biden and his administration! Or charge them with treason.	Neither	Intolerant
(b) Yes I have hope for your country. There are enough people who are sick of this.	Neither	Impolite
(c) Oh anyways the lefties are lying about everything relating to fixing the economy	Intolerant	Impolite
(d) How are you going to protect our Freedom? You Republicans don't even know how to keep the electricity on!	Intolerant	Neither
(e) FXCK THAT! NEVER GONNA HAPPEN IN AMERICA! Civil War will happen before that happens here! @LINK	Impolite	Neither
(f) When will this nincompoop leave the White House. He got 81 million votes? God help us!! #IllegitimatePresident	Both	Intolerant

Table 5: Examples of tweets illustrating discrepancies between human-assigned labels and classifier predictions for impoliteness and intolerance.

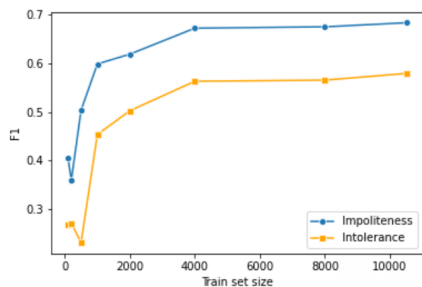


Figure 1: Test F1 results on impoliteness and intolerance detection, varying the number of training examples.

datasets of political incivility (Table 2).¹⁴ Concretely, we measured the extent to which performance declines in a cross-dataset setup compared to within-dataset training. We considered fixed random test sets (20%), finetuning a RoBERTa classifier in all cases. On average, applying our model to the other datasets resulted in lower precision (-25.3%) and higher recall (29%), reaching similar F1 results (-3.3%). We attribute the increased recall to the diversity of MUPID, where precision may be reduced due to data shift or incompatibility of the annotations. Inversely, we finetuned a model using the other datasets (19K examples overall) and applied it to MUPID test set. Compared to our results (Table 3), we observed lower precision (-11.5%), recall (-23.2%) and F1 (-18%). The reduction of recall reflects a failure to detect intolerant instances that are under-represented in the other datasets. See detailed results in Appendix B.

Impact of train set size. Figure 1 shows test F1 results while finetuning the RoBERTa classifier using increasing stratified subsets of the train set. As shown, impoliteness dominates intolerance detection results using as few as 1,000 training examples, again showing the greater semantic complex-

¹⁴The dataset due to Bianchi et al. (2022) is remote from ours for comparison purposes as it is focused on immigration.

ity involved in detecting uncivil substance vs. tone. While the improvement rate subsides past $\sim 4K$ labeled examples, the best results are obtained using the full dataset. We conjecture that similar to hate speech, further improvements may be achieved by extending the dataset, e.g., via methods of synthetic example generation (Wullach et al., 2021b; Hartvigsen et al., 2022).

5 From tweets to users: a large-scale evaluation

Automatic incivility detection may be used to identify and quantify political incivility at scale, addressing research questions of interest. Here, we introduce and examine the following questions: (i) Are certain users more inclined to post uncivil political content online? (ii) Do incivility levels vary by geopolitical region, specifically, across U.S. states? In both cases, we explore contextual factors that correlate with increased political incivility levels with respect to either impoliteness or intolerance.

To investigate these questions, we collected a corpus comprised of the twitting history of a large number of user accounts. Concretely, we randomly sampled users who authored tweets between July-Nov. 2022, whom we verified to be residents of the U.S. based on the location attribute of their profiles. For each user account, we retrieved the most recent (up to 200) tweets posted by them, discarding retweets and non-textual tweets, as well as tweets posted by overly active accounts suspected as bots.¹⁵ This resulted in a corpus of 16.3M tweets authored by 373K users. Out of those, 2.6M tweets by 230K users were classified as political, henceforth, *the corpus*. Finally, 17.6% of the political tweets were identified as impolite, 13.3% as intolerant, and 2.5% as both categories, accounting for overall incivility ratio of 28.4%. These proportions

¹⁵We removed accounts for which the tweet posting rate was higher than two standard deviations above the mean.

Variable	% Impolite	% Intolerant
User-level metrics (N=230K)		
# Followers	-0.109	-0.038
# Followees	-0.017	0.058
Tweets per day	0.068	0.091
% political tweets	0.237	0.498
Incivility among followees (N=1K, F=600k)		
% Impolite	0.135	0.236
% Intolerant	0.128	0.371

Table 6: Spearman’s correlations: the ratio of impolite/intolerant tweets vs. user-level metrics and the incivility ratios among the accounts followed. The table denotes the user sample size (N) and number of followees (F). All scores are significant (p -value < 0.001). Multivariate analysis gave similar results (Appendix C).

are similar to figures reported based on manual examination of a non-English political comments on Facebook—20% impolite and 10.8% intolerant comments (Rossini, 2022). Considering this distribution, we note the importance of detecting incivility both in terms of style and substance for achieving a comprehensive coverage of online hostility.

5.1 Political incivility at the user level

Our results indicate that some users are indeed more inclined to post uncivil content than others. As few as 7.3% of the users authored 50% of the uncivil posts in the corpus, and 20.6% of the users authored 80% of the uncivil posts. On the other hand, 43.7% of the users authored no uncivil post.

To explore the characteristics of incivility at user-level, we examined the associations between the share of impolite and intolerant tweets among one’s political tweets and other user-level metrics of interest, including network connectivity (number of followers and followees), activity level (average number of tweets per day), and the ratio of political tweets among the tweets posted by them. Table 6 reports our findings in terms of Spearman’s rank correlation scores. As shown, users who post intolerant and impolite political content are active, posting more tweets per day than other users. They also tend to have less followers—possibly, popular users refrain from controversial political language. Interestingly, a study of ‘hateful’ users similarly showed that they tweet more, follow other users more, but are less followed (Ribeiro et al., 2018). We find strong positive correlation between incivility and the share of political tweets posted by the user (Spearman’s correlation scores of 0.24 and 0.50 with respect to impoliteness and intolerance, respectively). That is, users who discuss political

topics more often—an indicator of increased political engagement (Vaccari and Valeriani, 2018), are more likely to use either intolerant or impolite language. This result echoes the suggestion that incivility may become normalized for those who discuss politics online more often (Hmielowski et al., 2014). As we observe similar trends for both types of incivility, our study suggests that public-level incivility, i.e., intolerance, may have also become normalized online among those who practice political talk often. Importantly, since our classifiers mainly focus on hostility between partisan groups and ideological camps (Table 4), our analyses and findings apply to this context.

In another analysis, we examine whether user-level incivility is correlated with incivility among the accounts that one follows. To address this question, we considered a random sample 1K users, and obtained the tweets posted by their followees within a 2-month period prior to the user retrieval date. Overall, we processed 8M tweets posted by 0.6M unique followees, quantifying the share of uncivil political tweets by those accounts. As detailed in Table 6 and in Appendix C, strong and significant correlations were found with respect to both types of incivility between users and the accounts that they follow. Thus, we observe a substantial degree of network homophily among users and followees who use political incivility online (see also Mathew et al. (2019)). This result implies that network information may provide meaningful context for political incivility detection, especially in those cases where indirect language is used (Ribeiro et al., 2018; Ghosh et al., 2023).

5.2 Incivility across geopolitical regions

Using our large sample of users, we further quantify and compare political incivility across geopolitical regions, namely, U.S. states. We identified relevant user accounts for this purpose, which specified state information (full state name, or its abbreviation) in the meta-data location field. Overall, 186K users in the corpus met this condition. The largest number of users were affiliated with the states of New York (23K), California (16K) and Texas (14K). The states with the least number of users were North Dakota (265), Wyoming (315), South Dakota (426), and Alaska (579). The median number of tweets per state was 2.2K, providing a sufficient sample size for statistical analysis.

For each state, we computed the average user-level proportion of impolite or intolerant tweets.

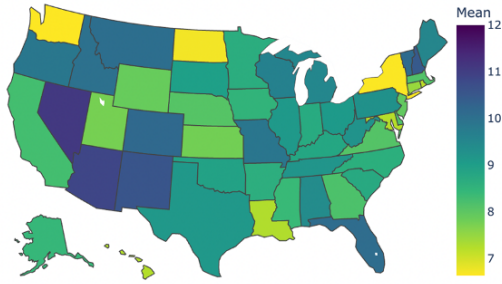


Figure 2: Average detected user-level political intolerance ratio per state (ranging between 7-12%).

Figure 2 presents a heat map illustrating the average age intolerance ratio across states. Also here, we observed aligned trends for both incivility types, obtaining similar results for impoliteness. As shown, some states demonstrate relatively low incivility rates (e.g., WA and NY) whereas others exhibit higher incivility rates (e.g., AZ and FL).

In light of these results, we conjectured that in ‘battleground states’, where the two camps are on par, there would be more hostility and toxicity in the political debate. To test this hypothesis, we contrasted the detected state-level average ratios of impolite and intolerant tweets against the differences between the percentage of votes for the Democratic and the Republican parties per state.¹⁶ The analysis confirmed our hypothesis, yielding significant Spearman’s rank correlation scores of -0.43 and -0.40 (p-value < 0.01), respectively. In words, this result suggests that political incivility tends to escalate in regions where electoral competition is intense, corresponding to a closer contest between the Democratic and Republican parties.

We note that rather than specify our results per state, we wish to highlight the contextual factors that may affect incivility rates at state-level. Our findings corroborate and align with existing literature of political communication. In particular, researchers previously showed that candidates and the media use more negative rhetoric in battleground states (Goldstein and Freedman, 2002); that citizens of battleground states engage more in politics on social media (Settle et al., 2016); and, that competitive districts feature higher levels of Twitter-based incivility (Vargo and Hopp, 2017). Our large-scale study is first to provide conclusive empirical evidence of increased multidimensional

political incivility by social media users in battleground states.

6 Conclusion

We presented MUPID, a dataset of political incivility annotated via crowd sourcing, distinguishing between dimensions related to style (impoliteness) and substance (intolerance). As discussed in detail, we refrained from term matching and from using available toxicity detection tools so as to diminish topical and lexical bias. Our experiments using finetuned language models and few-shot learners reached best F1 performances of 0.70 and 0.59 in identifying impolite and intolerance language, respectively. Our results and analyses suggest that finer semantic and social understanding is required for more accurately decoding incivility as perceived in political contexts, where this particularly holds for intolerant expressions. A large-scale study demonstrates the utility of our models for studying various aspects of political incivility. We find that users who are politically engaged, in that they post political content more often, are more inclined to use uncivil language, where as few as 20% of the users authored 80% of the uncivil tweets. We also track network homophily, showing that ‘uncivil users’ tend to follow other accounts with increased incivility. Analysing incivility at the aggregate level, we find that increased incivility is more prominent in battleground states.

Our dataset and models of multidimensional political incivility detection may support future research about the relationship between incivility and other contextual factors, e.g., user sociodemographics, as user traits such as age, gender, and education level, may be elicited given popular accounts that are followed by them (Lotan and Minkov, 2023). A temporal analysis may highlight the impact of political events on incivility levels.

We believe that political incivility detection would benefit from the modeling of relevant social context, such as conversation history (Ghosh et al., 2023) and the political events that the text refers to (Pujari and Goldwasser, 2021). Incorporating information about the user alongside the text authored by them may also help decode the text meaning (Pujari et al., 2024). Initial experiments, in which we conjoined user network embeddings with the text encoding showed improved prediction performance. We hope that researchers will benefit from our dataset in exploring similar directions.

¹⁶<https://www.cookpolitical.com/2020-national-popular-vote-tracker>

7 Limitations

This study applies to political incivility in the U.S., focusing on the Twitter network. Our dataset and models may be therefore limited geographically, temporally, and with respect to platform. In fact, soon after performing this research, the Twitter social network changed ownership and turned into X, where changes in its user base and political incivility levels might have followed. In general, however, we believe that much of the patterns captured in our dataset and models are general, and may transfer to other sites of social media and over time.

It is important to note that while we attend contextual factors of political incivility at user and geopolitical level, we acknowledge the potential significance of other contextual factors, e.g., the conversation history, and whether the discussion is held among like-minded users (Rossini, 2022). Exploring these aspects requires diverse methodological approaches, which are beyond the scope of the current paper.

Another limitation that is inherent to Twitter data concerns replicability, as accounts may be deleted or suspended and posts may be removed from the social network platform over time. This limitation applies to all Twitter datasets, which require tweet recovery via rehydration (Bianchi et al., 2022). We release our dataset, as well as our code and classification models to the research community to promote future research on this topic,¹⁷ and to allow comparison of our models with future models of political incivility detection.

8 Ethics statement

As the primary focus of this study is political incivility, crowd coders may have encountered texts characterized by an impolite style (e.g., foul language) or intolerant content (e.g., speech that discriminates against or excludes individuals based on their social and political characteristics). To mitigate potential harm to the crowd coders, we implemented several protective measures. First, we deliberately avoided providing coding examples that contained violent threats and extreme forms of incivility. Second, we ensured that all coding examples and tasks were derived from real-world political tweets, similar to those commonly encountered on social media platforms. Additionally, we allowed coders the flexibility to terminate their tasks at their

¹⁷The dataset and classification model are available on Hugging Face.

discretion. We further wish to clarify that political incivility is not considered to be a personal trait or a characteristic of a population by the authors. Considering that toxic political discourse may have become normalized among those who frequently engage in social media discussions, our study aims to distinguish between two distinct dimensions of such discourse within the framework of partisan competition. Finally, we clarify that the normative debate on online freedom of speech and its possible restrictions is beyond the scope of our manuscript. Rather, our study aims to provide a foundation for researchers to explore the underlying factors shaping political incivility, allowing for future studies to delve into its implications. This research was approved by our institutional review board.

Acknowledgements

We thank the reviewers for their useful comments. This research was funded by the Data Science and Research Center at the University of Haifa and by the Israeli Science Foundation, Grant no. 2671/22.

References

- Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901.
- 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):1531–1542.
- Sam Bestvater, Sono Shah, Gonzalo River, and Aaron Smith. 2022. Politics on twitter: One-third of tweets from us adults are political.
- Federico Bianchi, Stefanie Hills, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2022. “it’s not just hate”: A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.
- Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4):658–679.

- Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. [Developing a new classifier for automated identification of incivility in social media](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*.
- Flor Miriam Plaza Del Arco, Debora Nozza, and Dirk Hovy. 2023. [Respectful or toxic? using zero-shot learning with language models to detect hate speech](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. [Hate lingo: A target-based linguistic analysis of hate speech in social media](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Jeremy A Frimer, Harinder Aujla, Matthew Feinberg, Linda J Skitka, Karl Aquino, Johannes C Eichstaedt, and Robb Willer. 2023. [Incivility is rising among american politicians on Twitter](#). *Social Psychological and Personality Science*, 14(2):259–269.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Bryan T Gervais. 2014. [Following the news? reception of uncivil partisan media and the use of incivility in political expression](#). *Political Communication*, 31(4):564–583.
- Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. [Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Ken Goldstein and Paul Freedman. 2002. [Lessons learned: Campaign advertising in the 2000 elections](#). *Political Communication*, 19(1):5–28.
- Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. [Fake news on Twitter during the 2016 US presidential election](#). *Science*, 363(6425):374–378.
- Ilker Gül, Rémi Lebret, and Karl Aberer. 2024. [Stance detection on social media with fine-tuned large language models](#). *CoRR*, abs/2404.12171.
- Amy Gutmann and Dennis F Thompson. 2009. *Democracy and disagreement*. Harvard University Press.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *9th International Conference on Learning Representations (ICLR)*.
- Anushree Hede, Oshin Agarwal, Linda Lu, Diana C. Mutz, and Ani Nenkova. 2021. [From toxicity in online comments to incivility in American news: Proceed with caution](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. [A survey of methods for addressing class imbalance in deep-learning based natural language processing](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Jay D Hmielowski, Myiah J Hutchens, and Vincent J Cicchirillo. 2014. [Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming](#). *Information, Communication & Society*, 17(10):1196–1211.
- Jeffrey B Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2019. [Voteview: Congressional roll-call votes database](#). See <https://voteview.com/>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Nir Lotan and Einat Minkov. 2023. [Social world knowledge: Modeling and applications](#). *Plos one*, 18(7).
- Scott M Lundberg and Su-In Lee. 2017. [A Unified Approach to Interpreting Model Predictions](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. [Spread of hate speech in online social media](#). In *Proceedings of the 10th ACM conference on web science*, pages 173–182.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. [Hatexplain: A benchmark dataset for](#)

- explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- Ashley Muddiman. 2017. Personal and public levels of political incivility. *International Journal of Communication*, 11:21.
- Ashley Muddiman, Jamie Pond-Cobb, and Jamie E. Matson. 2020. Negativity bias or backlash: Interaction with civil and uncivil online political news content. *Communication Research*, 47(6):815–837.
- Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. [Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*.
- Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283.
- Rajkumar Pujari and Dan Goldwasser. 2021. [Understanding politics via contextualized discourse processing](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Rajkumar Pujari, Chengfei Wu, and Dan Goldwasser. 2024. “we demand justice!”: Towards social context grounding of political texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Ludovic Rheault, Erica Rayment, and Andreea Musulan. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1).
- Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.
- Patrícia Rossini. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Natalee Seely. 2018. Virtual vitriol: A comparative analysis of incivility within political news discussion forums. *Electronic News*, 12(1):42–61.
- Jaime E Settle, Robert M Bond, Lorenzo Coviello, Christopher J Fariss, James H Fowler, and Jason J Jones. 2016. From posting to voting: The effects of political competition on online political engagement. *Political Science Research and Methods*, 4(2):361–378.
- Rasmus Skytte. 2021. Dimensions of elite partisan polarization: Disentangling the effects of incivility and issue polarization. *British Journal of Political Science*, 51(4):1457–1475.
- Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Popa, and Olivier Parnet. 2016. A bad workman blames his tweets: The consequences of citizens’ uncivil Twitter use when interacting with party candidates: Incivility in interactions with candidates on Twitter. *Journal of Communication*, 66.
- Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. The dynamics of political incivility on Twitter. *SAGE Open*, 10(2).
- Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.
- Cristian Vaccari and Augusto Valeriani. 2018. Digital political talk and political participation: Comparing established and third wave democracies. *Sage Open*, 8(2).
- Jonathan Van’t Riet and Aart Van Stekelenburg. 2022. The effects of political incivility on political trust and political participation: A meta-analysis of experimental research. *Human Communication Research*, 48(2):203–229.
- 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):10–32.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. [Detection of abusive language: The problem of biased datasets](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Magdalena Wojcieszak, Andreu Casas, Xudong Yu, Jonathan Nagler, and Joshua A Tucker. 2022. Most users do not follow political elites on twitter; those who do show overwhelming preferences for ideological congruity. *Science advances*, 8(39):eabn9418.
- Magdalena Wojcieszak, Sjifra de Leeuw, Ericka Menchen-Trevino, Seungsu Lee, Ke M Huang-Isherwood, and Brian Weeks. 2023. No polarization from partisan news: Over-time evidence from trace data. *The International Journal of Press/Politics*, 28(3):601–626.
- Tomer Wullach, Amir Adler, and Einat Minkov. 2021a. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP*.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021b. [Towards hate speech detection at large via deep generative modeling](#). *IEEE Internet Comput.*, 25(2):48–57.

Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. [When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers](#). In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Avishai Zagoury, Einat Minkov, Idan Szpektor, and William W. Cohen. 2021. [What’s the best place for an AI conference, vancouver or _____: Why completing comparative questions is difficult](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*.

A Instructions and interfaces for the crowd workers and the GPT prompt

Figure 3 presents the code book presented to the crowd workers, and Figure 4 demonstrates the training phase which workers had to complete in order to get qualified to work on our task. As shown in the screenshot, following the completion of the training phase, the candidate worker was presented with explanations about their labeling mistakes. In instructing the GPT model to label the test examples, we used the prompt shown in Figure 5.

B Cross-dataset evaluation results

Table 7 includes detailed intra- and cross-dataset evaluation results.

	Train	Test	Precision	Recall	F1
MUPID → Other datasets:					
I	TH	TH	0.677	0.543	0.604
C	MUPID	TH	0.542	0.847	0.661
		Δ	-19.9%	56.0%	9.4%
I	RH	RH	0.845	0.672	0.749
C	MUPID	RH	0.547	0.831	0.66
		Δ	-35.3%	23.6%	-11.9%
I	DA	DA	0.871	0.725	0.791
C	MUPID	DA	0.692	0.779	0.733
		Δ	-20.6%	7.4%	-7.3%
		Average Δ:	-25.3%	29.0%	-3.3%
Other datasets → MUPID:					
I	MUPID	MUPID	0.765	0.707	0.735
C	All	MUPID	0.677	0.543	0.603
		Δ	-11.5%	-23.2%	-18.0%

Table 7: Detailed cross-dataset evaluation results: Intra-(I) vs. cross-dataset (C) experiments. The table uses acronyms: TH (Theocharis et al., 2020), RH (Rheault et al., 2019), DA (Davidson et al., 2020).

Variable	Odds ratio	Std.Error
IMPOLITE		
# Followers	1.000000	1
# Followees	0.999992	1.000001
Tweets per day	1.008036	1.000401
% Political tweets	1.589433	1.020808
INTOLERANT		
# Followers	1	1
# Followees	1.00001	1.000001
Tweets per day	1.008002	1.000356
% Political tweets	5.176365	1.018723

Table 8: Multivariate beta regression results of user-level characteristics as explaining factors of the share of impolite and intolerant tweets out of their political tweets. The sample size is 230K users, and all the results are significant at p-value < 0.001.

C Multi-variate analyses of user-level incivility

This section include multi-variate analysis results, showing similar trends to our results measured in terms of Spearman’s correlation, reported in Table 6.

We modeled multivariate beta regressions to examine the associations between the share of impolite and intolerant tweets out of users’ political tweets and other user-level characteristics, including their number of followers, number of followees (i.e., accounts followed by a given user), average tweets per day, and the share of political tweets out of the total texts by a given user. The correlates with respect to the ratio of impolite and intolerant tweets are presented in Tables 8. We use odds ratio (OR) to interpret the results more intuitively. The results show, for example, a positive relationship between the share of impoliteness and tweets per day (OR = 1.008): for a one-unit increase in a user’s tweets per day, the odds of observing a higher share of impolite tweets increase by 0.80%. Focusing on the share of political tweets as a predictor, the results show that a movement from its minimum value (0) to its maximum value (1) is associated with a 59% increase in the odds of observing a higher share of impolite tweets (OR = 1.59). We also observe that a greater share of political tweets is associated with a higher ratio of intolerant tweets, to a greater extent (OR = 5.17). Note that while there is a very small change in impoliteness or intolerance ratio with the increase of a single follower or followee (OR is roughly 1), this effect is statistically significant.

We also examined whether posting uncivil tweets is correlated with exposure to incivility by

Rules and Tips

1. What makes a tweet **uncivil**:

The tweet contains foul language or a harsh tone toward other people or their ideas and actions. It can also include harmful or discriminatory intent toward people or groups based on gender, race, ethnicity, political views, etc.

2. **Uncivil tweets can be categorized into three sub-dimensions**: (A) impoliteness, (B) intolerance, or (C) both.

A. **Impoliteness**: the tweet contains insults, foul language, harsh tone, name-calling, vulgarity, an accusation of lying, or aspersion toward other people or their ideas and actions.

B. **Intolerance**: the tweet contains expressions that **derogate or undermine particular groups** due to social, political, sexual, ethnic, or cultural features. The tweet can contain threats of physical or emotional harm to others, or the silencing or denial of rights of people and groups (e.g., minorities, political groups, etc.).

C. **Both**: the tweet contains both of the above sub-dimensions.

3. An intolerant tweet (sub-dimension B) **does not** necessarily have an impolite style (sub-dimension A), and vice-versa.

4. Disagreements with another person or idea are **not** considered uncivil automatically. When these disagreements contain foul language or intolerant expressions, only then the tweet will be categorized as uncivil.

5. Notice that a tweet may contain criticism (even harsh criticism) toward public figures. However, criticism does not automatically make the tweet intolerant. In order to be intolerant, the tweet should explicitly offend/derogate entire groups (e.g., minorities, liberals, Republicans, women, LGBT, etc.)


Figure 3: The code book presented to the crowd workers

one's followees network (i.e., the accounts that the user follows). We calculated users' potential exposure to incivility as the share of impolite and intolerant tweets in their followees' network, i.e., the number of uncivil tweets posted by their followees divided by the total number of political tweets of these followees. We modeled the same beta regressions as above, this time adding considering the share of impolite and intolerant tweets in one's network as predictors. In the case of impoliteness, it is indicated that the more users are potentially exposed to impoliteness in their network, the higher is the share of their impolite tweets (OR = 1.03, a 3% increase, p-value < 0.001). Similar findings are observed in the case of intolerance (OR = 1.06, a 6% increase, p-value < 0.001). While we cannot establish causality due to the cross-sectional nature of the data, we encourage scholars to further investigate these initial conclusions that uncivil users follow others who behave similarly.

You were right in 2 out of 6 questions.





Correct answers in green & incorrect ones in red color.

Please review the correct answers and their detailed explanations:




USER @User · Jun 27

The government has ruined millions of lives and livelihoods by disrupting commerce, destroying small businesses and hindering individuals from going about their daily lives. That's what the truckers are protesting, you rank imbecile. People like you need a damn good kicking.







This is an uncivil tweet that belongs to the impoliteness category. Although the beginning of the tweet is critical of the government in a relatively civil manner, the tweet ends with offensive language, including direct insults.



USER @User · Jun 27

It would be nice to see the left fight for the basic human rights of everyone like they do for abortions.



This is a neutral tweet that is considered Civil. For a tweet to be labeled uncivil, it must be unambiguously offensive or contain derogative language. In this case, the tweet criticizes liberals in a relatively neutral way.

Figure 4: Annotator training: each candidate worker was asked to label 6 example tweets. We provided feedback about the labeling accuracy, and explanations about the labeling mistakes. Workers who obtained 4 out of 6 or higher accuracy in their responses got qualified to work on our task.

You are an advanced Political Impoliteness Classification System, specialized in detecting impolite language and behavior in political discourse, particularly within the context of United States politics and its specific terminology.

Instruction:

Determine whether the provided tweet displays impoliteness.

Impoliteness is characterized by insults, foul language, harsh tone, name-calling, vulgarity, an accusation of lying, or aspersion toward other people or their ideas and actions.

Please provide your judgment in the following JSON format: {"impoliteness": "Yes" or "No"}

Example Evaluations:

Tweet: "All hell has broken loose under the leadership of the senile old man. And now due to his weakness we will see him take us to WWIII. Young people voted for this crap."

Your JSON response: {"impoliteness": "Yes"}

Tweet: "And what's it called when Hillary and the dems arranged illegal surveillance against the POTUS? spying on the Whitehouse servers? Hillary and the dems ARE enemies, foreign AND domestic."

Your JSON response: {"impoliteness": "No"}

Tweet: "@USER just passed a trillion dollar infrastructure bill for Biden with no wall funding. How long do Republicans believe you can keep pushing this line? You never intended to secure the border."

Your JSON response: {"impoliteness": "No"}

Tweet: {x}

Your JSON response:

Figure 5: The prompt provided to the GPT-3.5-instruct model for impoliteness classification. A similar prompt was provided for intolerance classification. The format of the prompt follows common practice in instructing GPT-instruct and similar models to perform specific classification tasks.