Multidimensional Political Incivility on Twitter: Detection and Findings

Anonymous ACL submission

Abstract

Hostile and uncivil political discourse online may negatively affect Democratic processes. In this work, we consider the task of *political incivility* detection. Unlike previous attempts, we utilize a multidimensional perspective of political incivility, differentiating between impoliteness–which is sometimes acceptable–and intolerance–which inherently violates the Democratic norms. We evaluate state-of-the-art classifiers on this task using MUPID, a large multidimensional political incivility dataset of 13K tweets, which we collected and annotated by means of crowd sourcing. Our results and analyses illustrate the challenges involves in this task. In particular, we observe that intolerance is often expressed using implicit language, that requires higher-level semantic understanding. In addition, we apply political incivility detection at large-scale, exploring the distribution of uncivil content across individual users and across U.S. states. Our findings align and extend existing theories of Political Science and Communication.[1]

1 Introduction

A civil discourse between political groups is considered a fundamental condition for a healthy democracy (Gutmann and Thompson, 2009). The rise of social media has been argued to intensify disrespectful and hostile online political discourse (Coe et al., 2014; Frimer et al., 2023). This trend has multiple negative consequences: it fosters polarization between rival political groups, decreases trust in political institutions, and may disengage citizens from being politically involved (Muddiman et al., 2020; Skytte, 2021; Van't Riet and Van Stekelenburg, 2022). Considering these concerns, several research works have attempted to detect, quantify and characterise political incivility in discussion groups and social media platforms (ElSherief et al., 2018; Davidson et al., 2020; Theocharis et al., 2020; Frimer et al., 2023). These efforts offered however a coarse definition of incivility, placing emphasis on the tone of the conversation.

This work follows recent theories of political communication that view political incivility as a multidimensional concept (Muddiman, 2017; Rossini, 2022). The first dimension is *personal-level incivility (impoliteness)*, pertaining to a violation of interpersonal norms. Impolite speech may contain foul language, harsh tone, 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)* refers to violations of norms related to the democratic process, such as pluralism and deliberation. It pertains to exclusionary speech, silencing or denying the rights of social and political groups (e.g., “You are anti Americans, .. you should all leave and find a communist country”). Making a distinction between tone and substance, it is argued that heated political talk should not be necessarily dismissed due to impoliteness, as opposed to intolerant discourse, which describes democratically threatening behaviors. Hence, scholars interested in understanding the extent to which digital platforms threaten democratic values should focus on expressions of intolerance (Papacharissi, 2004; Rossini, 2022).

To the best of our knowledge, this work presents the first empirical attempt to detect political incivility at this resolution. A main contribution of our work is the construction of a large dataset of 13K political tweets, labeled by multidimensional incivility. The data collection process involved diverse sampling strategies, aiming at capturing both incivility types while avoiding lexical biases. We make this resource available to the research community.

Using our dataset, we adapt and evaluate a vari-

[1] The camera-ready version will include links to our dataset and model on Hugging Face. Meanwhile, an anonymous GitHub link is provided.

ety of state-of-the-art language models on the task of multi-label incivility detection. Our empirical investigation highlights the differences between impoliteness and intolerance. Both concepts are somewhat subjective and challenging for automatic language processing, yet intolerant utterances tend to be implicit and lexically ambiguous, calling for higher-level semantic and social understanding.

Another contribution of this work is a large scale study, where we examine the prevalence of incivility among more than 200K users who posted political content on Twitter.[2] Concretely, we explore whether (i) some individual users are more inclined to use impolite and intolerant language than others, and whether (ii) there are differences in incivility among geopolitical subpopulations, specifically, across states. Our findings corroborate and complement existing theories of political science.

2 Related work

There exists theoretical and empirical ambiguity in the literature regarding the definition of political incivility. According to a recent survey, there is considerable overlap among instances defined as uncivil, offensive, or toxic, where the term incivility is most frequently used by social scientists (Pachinger et al., 2023). Focusing on political incivility, some researchers framed it in terms of impolite speech (Theocharis et al., 2016; Seely, 2018), whereas others defined it as either impoliteness, intolerance or hate speech (Davidson et al., 2020; Theocharis et al., 2020). Relevant empirical studies addressed incivility detection as a binary classification problem (Davidson et al., 2020; Theocharis et al., 2020; Rheault et al., 2019). Following recent theories of Political Communication, this work considers political incivility as a multidimensional concept, defining uncivil language as either impolite or intolerant, or both. Rossini (2022) previously explored multidimensional incivility by manually coding a large sample of political comments posted by Facebook users in Brazil.[3] Here, we present a relevant dataset that includes political tweets of the U.S. context. We evaluate and apply incivility detection at scale, demonstrating the potential of this approach for studying questions of interest to political communication research.

We find that there exist relatively few works by researchers of natural language processing that examine incivility detection in the context of party competition in the U.S., e.g., (Hede et al., 2021). Public-level incivility (intolerance) is a broad concept, which pertains to exclusionary speech against groups of opposing beliefs, as well as against social minorities, aka *hate speech*. The task of hate speech detection is well-studied, yet far from solved. In particular, researchers have recently pointed out the challenges involved in the detection of implicit hate speech, where the underlying toxic intention is encoded via target-specific semantic frames rather than by foul language (ElSherief et al., 2021; Hartvigsen et al., 2022). In our work, we show that political intolerance is often conveyed in an implied manner. Thus, with respect to political incivility detection, we extend existing efforts to include intolerance that is not necessarily impolite. In the context of hate speech, we study another type of political intolerance, that is directed at opposing political groups rather than at social minorities.

3 MUPID: a Multidimensional Political Incivility Dataset

3.1 Data sampling strategy

Even though political incivility is not rare, the inspection of random tweets would yield a low ratio of relevant examples at high annotation cost. We exploited multiple network- and content-based cues, aiming to obtain a diverse sample of relevant tweets of the target classes while avoiding lexical and other biases (Wiegand et al., 2019).

Initial sampling. Our initial pool of tweets includes tweets by users who follow multiple disputable political accounts, assuming that those users may be more inclined to use uncivil language in political contexts (Gervais, 2014). Concretely, we referred to 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 ideologically extreme (Lewis et al., 2019).[4]We selected the most biased accounts per category, balanced over conservative and liberal orientation, based on bias scores specified by those sources.[5] Having identified users who followed at

[2] This research was performed before Twitter changed its name to X.

[3] Their annotated dataset is not publicly available.

[4] https://voteview.com/data

[5] We selected 20 accounts per source and orientation, except the fake news category, which includes only 9 accounts of each orientation.

least two of those biased accounts, while maintaining a balance between users of conservative and liberal orientation, we retrieved the (200) most recent tweets posted by them as of December 2021. This yielded 885K tweets authored by 15.8K users.

Identifying political tweets. As we study incivility in political contexts, it was necessary to identify tweets of topical relevance. We trained a dedicated classifier, exploiting existing data resources for this purpose. Specifically, we sampled 12.5K tweets from a large collection of tweets that concern political topics discussed frequently by either Republicans (e.g., the U.S. federal budget), Democrats (e.g., marriage equality, raising the minimum wage), or both (e.g., the presidential campaign) (Barberá et al., 2015). Additional 3.5K political posts were extracted from the social media accounts of U.S. politicians.[6] As counter examples, we considered random tweets by U.S. users,[7] constructing a balanced dataset of 32K political and counter 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 tuned classifier to our pool of sampled tweets, we set a high threshold (0.96) over the classifier confidence scores, targeting high precision in identifying tweets as political. Overall, 82K (9.3%) of processed tweets were predicted to be political. A manual examination of 300 random tweets by a graduate student of Communication indicated on prediction precision of 0.91 (273/300).

Sampling tweets for annotation. In order to focus the human annotation effort on tweets that demonstrate incivility, we followed several additional sampling heuristics. First, similar to previous works (Theocharis et al., 2020; Hede et al., 2021), we utilized the pretrained Jigsaw Perspective tool[8] to identify toxic tweets, sampling roughly 2K tweets that received high scores on the categories of 'abusive language and slurs', 'inflammatory comments' and 'attacks on the author'. In addition, following insights by which hateful user accounts tend to be new, and more active than average (Ribeiro et al., 2018), we sampled 2K accounts which were created up to two months prior to sampling date, or have posted more than one tweet a day on average since their creation date. Finally, we sampled 4K political tweets uniformly at random. Throughout the annotation process, we maintained 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. In order to obtain more examples of intolerant tweets, we followed an active labeling paradigm (Tong and Koller, 2001), where we employed a classifier of intolerance detection trained using the examples labeled thus far to identify additional tweets that were likely to be intolerant within our 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 among those tweets was similar, yet the ratio of intolerant tweets has tripled.

3.2 Annotation procedure and results

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 turn to crowd sourcing, using the platform of Amazon Mechanical Turk.[9] In order to elicit labels of high-quality, we required the workers to undergo dedicated training and quality testing, having the final labels determined based on the judgements of multiple qualified workers. We acknowledge that human judgement on this semantic task is subjective, being affected by one's cultural background, beliefs, and political stance. The inter-annotator disagreement ratios reported below gives an indication for the semantic complexity and subjectivity of the target concepts.

Procedure. Given each tweet, several independent workers were asked to determine whether it was impolite, intolerant, neither, or both.[10] 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. Figure 1 shows the annotation interface.

We required the workers to be highly qualified (having previously completed at least 100 micro-tasks with approval rate above 98%). We also restricted the task to residents of the U.S., to assure fluency in English and familiarity with U.S. politics.

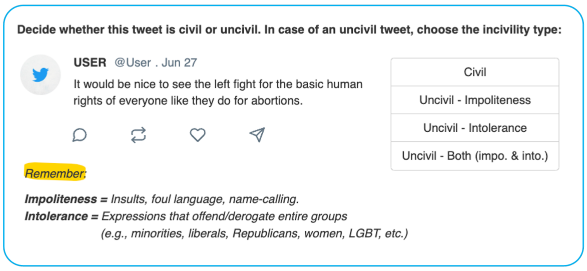
[6] www.kaggle.com/datasets/crowdflower/political-social-media-posts

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

[8] https://www.perspectiveapi.com/

[9] www.mturk.com/

[10] We specified the category of 'both' in order to raise annotator awareness of this possibility.



Decide whether this tweet is civil or uncivil. In case of an uncivil tweet, choose the incivility type:

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.

Remember:

Impoliteness = *Insults, foul language, name-calling.*
Intolerance = *Expressions that offend/derogate entire groups*
(e.g., minorities, liberals, Republicans, women, LGBT, etc.)

Figure 1: Annotator interface: the workers were asked to label tweets as impolite, intolerant, neither or both.

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 which were presented to the annotators as part of their training.

Candidate workers were further asked to undergo a qualification phase, in which they labeled six selected tweets. Whoever labeled a majority of the tweets correctly got qualified to work on our task, as well as received detailed feedback for their mistakes. During annotation, we included control questions (2 out of 15 tweets in each micro-task) that we expected the workers to do well on. In case that the worker failed to label the control tweets correctly, we rejected their annotations, 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, where fair pay positively affect 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.

Inter-Annotator Agreement. Each tweet was labeled by 3-5 annotators, so as to break label ties. We discarded examples for which a label could not be determined based on majority voting.[11] Comparing the labels assigned to every individual tweet by random worker pairs resulted in Fleiss’ kappa scores of 0.63 and 0.54, indicating on ‘substantial’ and ‘moderate’ agreement with respect to the category of impoliteness and intolerance, respectively. This suggests that intolerance is more subjective and subtle compared to impoliteness. Accordingly, a larger number of annotations was acquired on average for the intolerance category (3.21 vs. 3.07). We further assessed the quality of the final majority labels against the judgement of a Communication scholar for 300 random tweets drawn from our dataset. Fliess’ kappa scores both indicated on ‘substantial agreement’, measuring 0.57 and 0.61 on impoliteness and intolerance, respectively. Considering the subset of examples for which the workers tended to agree on, with high majority of 70% or more, showed substantially higher agreement with the expert on the impoliteness category compared with intolerance, measuring 0.78 vs. 0.69, respectively. Again, this suggests that the notion of political intolerance is more semantically subtle.

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.

3.3 Dataset statistics

Overall, MUPID (a MUltidimensional Political Incivility Dataset) includes 13.1K labeled tweets. As detailed in Table 2, a large proportion of the tweets (42.3%) are labeled as uncivil, including 3.6k impolite and 2.3K intolerant tweets. In comparison, existing related datasets use binary annotations, and include substantially fewer (0.6-1.2K) incivility examples. We make the dataset available for researchers. For each example, we specify its sampling method. Importantly, the examples obtained via active sampling were all allocated to the training set in our experiments in order to avoid evaluation bias.

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 coarse 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

[11] An odd number of annotations is not guaranteed to break ties for more than 2 categories.

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-related 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.[12] 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 target class. The table includes also the results of binary classification experiments, considering incivility as a unified coarse concept. As shown, identifying incivility at coarse-level 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 Google 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 shown 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 means that intolerance is typically not conveyed using general foul language.

Finally, considering that the Generative Pre-trained Transformer (GPT) model has been applied to related tasks such as hate speech detection (Del Arco et al., 2023), Table 3 displays our results of attempting few-shot incivility prediction using GPT-3.5.[13] In this case, for each target category, we prompted the model with a definition of the task and category, as well as with the same labeled examples presented to the human workers (Appendix A). As shown in the table, this method yielded relatively high performance given a handful of examples. However, it wasn’t among the best-performing methods. Possibly, improvements may be achieved via prompt engineering, but this is non-trivial and out of the scope of our work. Nevertheless, we observe that the results using GPT-3.5 are aligned with the general trends, achieving significantly lower performance in detecting intolerant vs. impolite speech (F1 of 0.44 vs. 0.58).

Overall, the finetuned DeBERTa and RoBERTa achieved the best performance. Henceforth, we will consider RoBERTa as our classifier of choice.

Impoliteness vs. intolerance. We applied Shapley analysis (Lundberg and Lee, 2017)[14] to our training set to identify unigrams predictive of impoliteness or intolerance. As shown in Table 4, impolite speech 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, we observe that the word types associated with political intolerance often refer to a political camp, e.g., ‘republicans’, or ‘liberals’. Unlike slur words, the sentiment of such terms is subjective or 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%). This suggests that high-level semantic and contextual understanding is needed to detect intolerance.

Table 5 includes examples of misclassified tweets, noting their assigned and predicted labels. We indeed observe cases in which the model missed the presence of intolerance due to its implied expression (examples (c) and (d)), e.g., “you Republicans don’t even know how to keep the electricity on!”. On the other hand, the model was sometimes misled by lexical cues, demonstrating the gap between lexical-level and semantic understanding; 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

[12] https://huggingface.co/

[13] GPT-3.5-turbo-instruct, see https://platform.openai.com/docs/models

[14] https://github.com/slundberg/shap

	Impolite				Intolerant				Coarse				
Classifier	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	Mac.F1
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
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
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

Table 3: Multi-label and binary prediction results.

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
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)

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

and his administration! Or charge them with treason" was justifiably classified as intolerant. Overall, this demonstrates the semantic and contextual challenges involved in identifying political intolerance.

Cross-dataset evaluation. Considering potential lack of generalization due to dataset bias (Wiegand et al., 2019), we assessed learning generalization using MUPID, applying our model of coarse incivility prediction to other datasets (Table 2). Concretely, we measured the extent to which performance declines in this 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 high diversity of MUPID. The reduced precision may be due to data shift, or incompatibility of the annotations. In a similar experiment, we finedtuned 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. More detailed results are included in Appendix B.

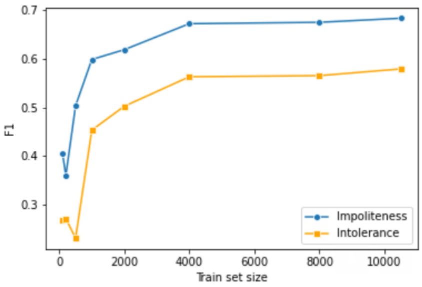


Figure 2: Test F1 results on impoliteness and intolerance detection, varying the number of training examples.

Impact of train set size. Figure 2 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. This is yet another evidence of the greater semantic complexity involved in political intolerance detection. While the improvement rate subsides past ∼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, up to magnitudes of order, via methods of synthetic example generation (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, as well as address 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 subpopulation, specifically, across U.S. states? 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

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	Impolite	Intolerant

Table 5: Evaluation of predicted examples.

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).

overly active accounts suspected as bots.[15] This resulted in a corpus of 16.3M tweets authored by 373K users. Out of those, 2.57M 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 an overall incivility ratio of 28.4%. These proportions are in the same ballpark of figures reported previously based on a manual examination of a non-English political comments on Facebook–20% impolite and 10.8% intolerant comments (Rossini, 2022).[16]

5.1 Political incivility at 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 Spearmans's 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, being more politically engaged (Vaccari and Valeriani, 2018), are more likely to use 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).

In another analysis, we examine whether user-level incivility is correlated with incivility among the accounts that one follows. To bound computation cost, we considered a random sample 1K users. We 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 account followed, detecting and quantifying the share of impolite and intolerant political tweets posted by those accounts. As detailed in Table 6 and in Appendix C, strong and significant correlations are found between the users' levels of impoliteness and intolerance and the political incivility levels of the accounts that they follow. This result suggests that network information may be informative for political incivility detection (Ribeiro et al., 2018).

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

[16] Their work pertains to Facebook and the Web in Brazil.

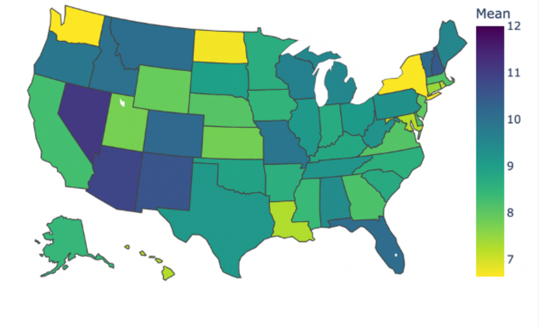


Figure 3: Average detected user-level political intolerance ratio per state (ranging between 7-12%).

5.2 Political incivility across U.S. states

To quantify and compare political incivility across U.S. states, we attended user accounts that 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 2,216, providing a sufficient sample size for statistical analysis.

For each state, we computed the average user-level proportion of impolite or intolerant tweets. Figure 3 presents a heat map showcasing the average intolerance ratio across states. Similar trends were observed for impoliteness. As shown, some states demonstrate low incivility rates (e.g., WA and NY) whereas others exhibit high incivility rates (e.g., AZ and FL). Given these findings, 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.[17] The analysis confirmed our hypothesis, yielding significant Spearman’s correlation scores of -0.43 and -0.40 (p-value $<$ 0.01), respectively. In words, this result suggests that higher levels of political incivility in a particular state correspond to a closer contest between the two main political parties. These findings align with existing literature. Researchers previously showed that candidates and the media use more negative rhetorics 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 work is first to provide empirical evidence of increased multidimensional political incivility by social media users in battleground states.

6 Conclusion

Considering its negative implications, it is desired to identify and address political incivility online. Following recent theories by Political Science researchers, we distinguish between impolite discourse, which may be acceptable in heated discussions, and intolerance, that violates Democratic norms. Framing political incivility detection as a multidimensional classification task, we presented MUPID, a large dataset annotated via crowd sourcing at this resolution. Our experiments using fine-tuned language models and other popular methods reached best F1 performances of 0.70 and 0.59 in identifying impolite and intolerance language, respectively, indicating that is a challenging task. Our results and analyses suggest that better semantic and social understanding is required for more accurately decoding incivility as perceived in political contexts, where this particularly holds for intolerant expressions. Based on a large-scale study of incivility at user-level, we assert that certain individuals are more inclined to political incivility. Those users seem to be more engaged, posting political content more often, and tend to follow other accounts with increased incivility. Analysing incivility in aggregate across states, we found that increased incivility is more prominent in battleground states. These results align with and extend existing literature, demonstrating the potential of studying political incivility using automated models.

Future research may continue to explore the relationships between incivility and other factors, e.g., sociodemographics and political stance. Political incivility prediction may be improved by modeling user and conversation contexts (Ghosh et al., 2024). We hope researchers will propose methods for moderating political incivility (Tekiroğlu et al., 2020), mainly, its more severe form of intolerance.

[17] 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. While we targeted the detection of political intolerance as a broad concept, the tweets annotated as intolerant in our dataset mostly undermine or silence partisan political groups (e.g., 'republicans', 'democrats', or 'liberals'). It is possible that political intolerance in its bipartisan context is inherently more prevalent in Twitter.

Methods-wise, we applied political tweets as a preliminary step. This approach was mainly motivated by computational considerations, however joint processing may be desired in other circumstances. Content-wise, our dataset and models may be 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 our approach and models may be applied however to track incivility in other sites of social media and over time. It is possible that the dataset and models may be enhanced using methods such as data generation (Hartvigsen et al., 2022).

There are several other limitations that are inherent to the task. Incivility is a subjective concept, as reflected by moderate agreement rates. As we point out in the work, it involves fine semantics, where intolerance typically takes an implicit form. We believe that the modeling of relevant social knowledge, e.g., the political stance of the author of the tweet and their network characteristics, could complement content-based methods.

8 Ethics statement

This research was approved by our institutional review board. Unlike hate speech, which targets and threatens minorities, partisan intolerance negatively impacts democratic processes, but may not be as offensive or threatening at a personal level. We therefore assess the potential harm of exposure to the political incivility examples included in this paper, and in our dataset, to be low. The labeling process involved subjective judgement, noting incivility as perceived by human workers, rather than the meaning as intended by the text authors. We release our code and dataset to the research community in compliance with Twitter terms to promote future research on this topic.

References

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.

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. 2024. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

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.

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.

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, 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.

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.

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.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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, 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.

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)*.

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

Figure 4 presents the code book presented to the crowd workers, and Figure 5 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 6.

B Cross-dataset evaluation results

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

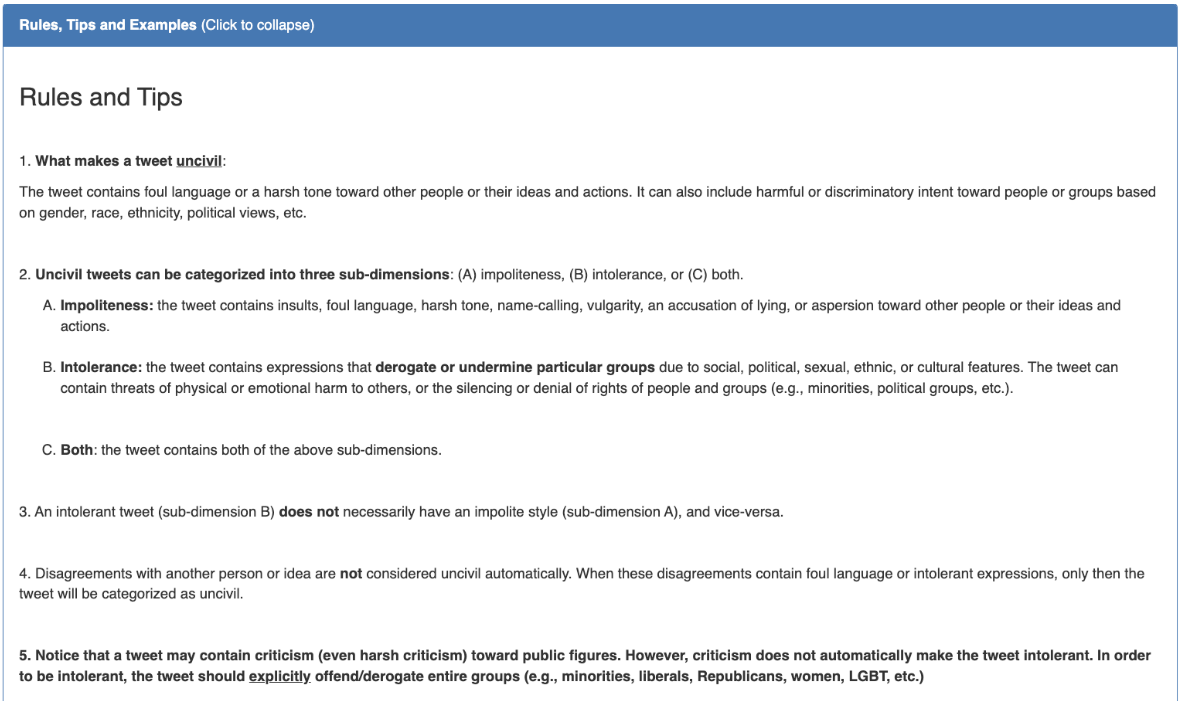
	Train	Test	Precision	Recall	F1
MUPID → Other datasets:					
I	TH	TH	0.677	0.543	0.604
C	MUPID	TH	0.542	0.847	0.661
		Δ	-19.9%	56.0%	9.4%
I	RH	RH	0.845	0.672	0.749
C	MUPID	RH	0.547	0.831	0.66
		Δ	-35.3%	23.6%	-11.9%
I	DA	DA	0.871	0.725	0.791
C	MUPID	DA	0.692	0.779	0.733
		Δ	-20.6%	7.4%	-7.3%
		Average 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).

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



Rules, Tips and Examples (Click to collapse)

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.)

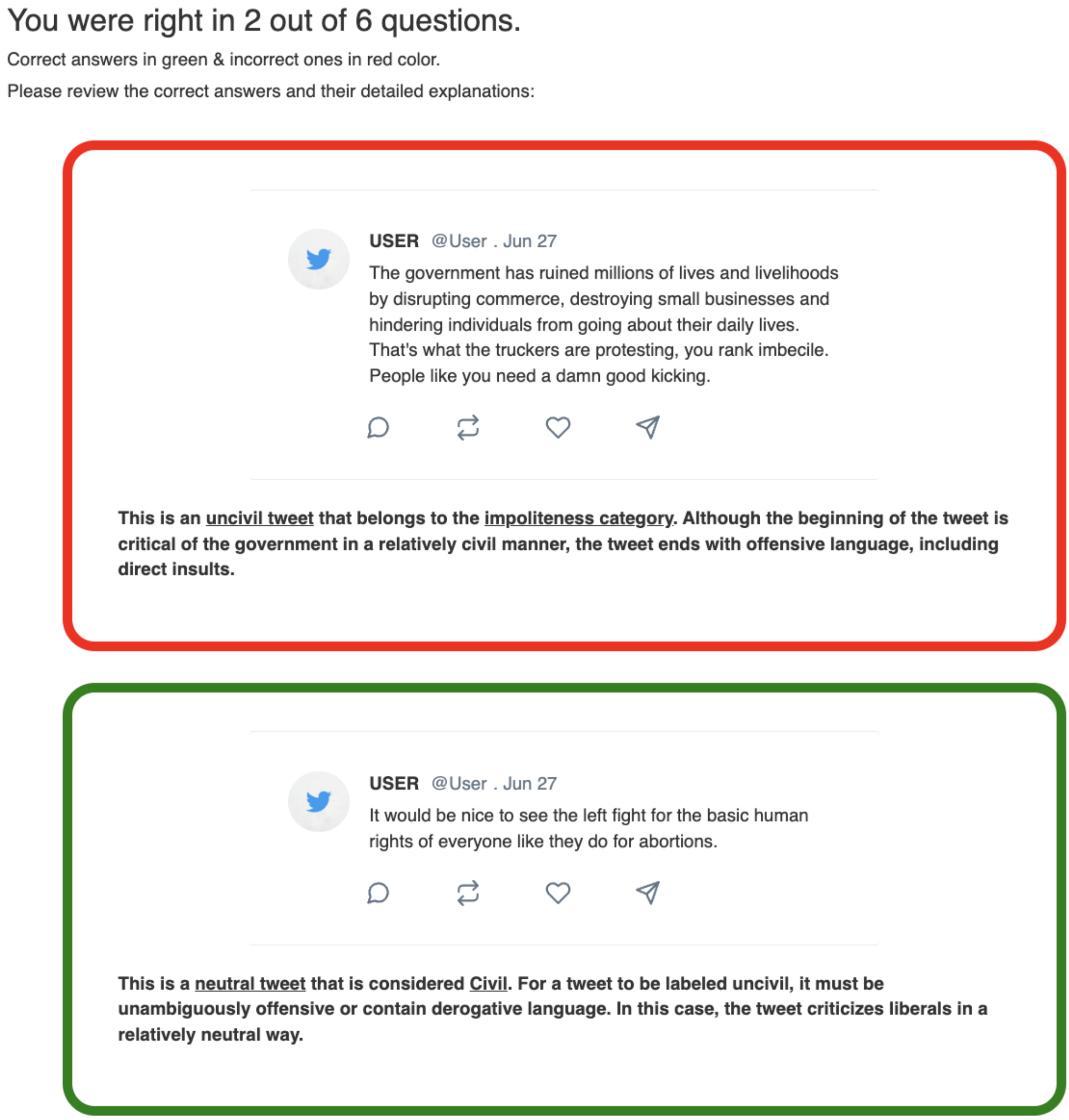
Figure 4: The code book presented to the crowd workers

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 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.

Variable	Odds ratio	Std.Error	Significance
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.



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.

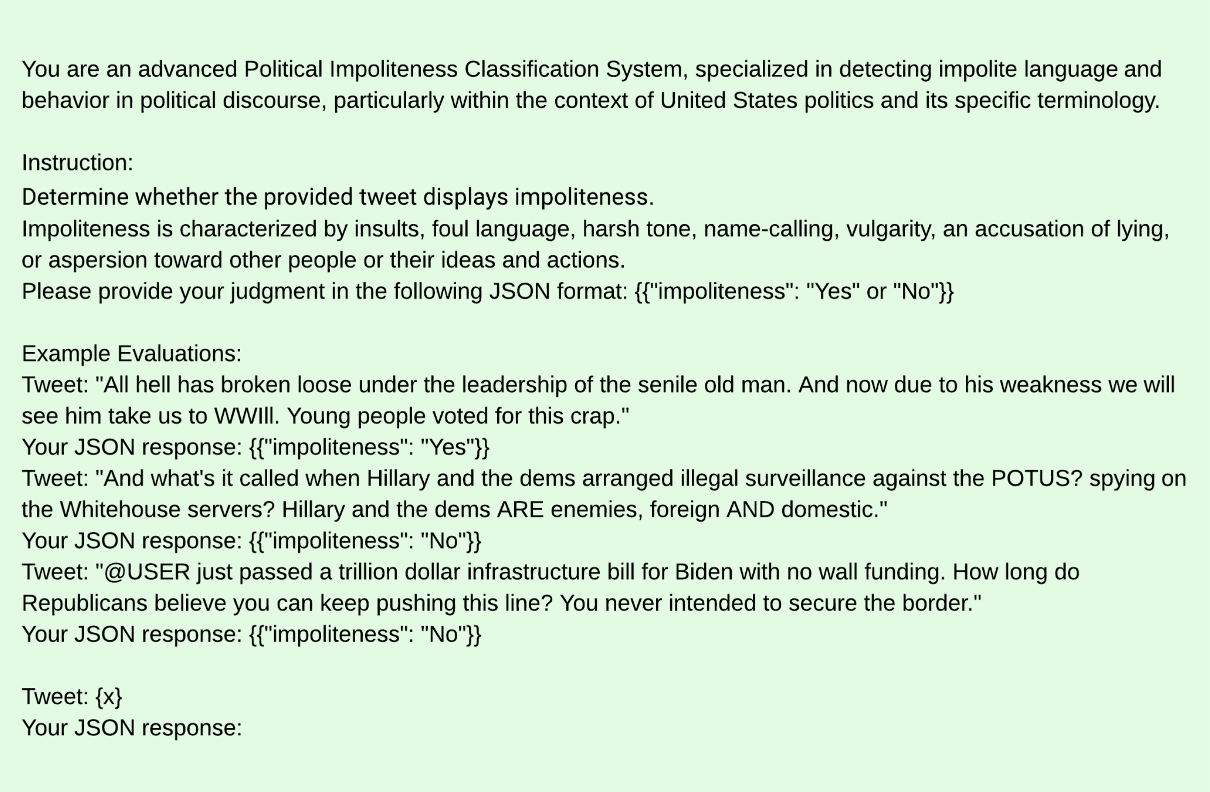
This is an uncivil tweet that belongs to the impoliteness category. Although the beginning of the tweet is critical of the government in a relatively civil manner, the tweet ends with offensive language, including direct insults.

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 derogatory language. In this case, the tweet criticizes liberals in a relatively neutral way.

Figure 5: 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 WWIll. 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 6: 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.