Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Hate Cannot Drive out Hate: Forecasting Conversation Incivility following Replies to Hate Speech

Xinchen Yu[1*], **Eduardo Blanco**[1], **Lingzi Hong**[2]

[1]Department of Computer Science. University of Arizona
[2]Department of Information Science. University of North Texas
{xinchenyu, eduardoblanco}@arizona.edu, lingzi.hong@unt.edu

Abstract

User-generated replies to hate speech are promising means to combat hatred, but questions about whether they can stop incivility in follow-up conversations linger. We argue that effective replies stop incivility from emerging in follow-up conversations—replies that elicit more incivility are counterproductive. This study introduces the task of predicting the incivility of conversations following replies to hate speech. We first propose a metric to measure conversation incivility based on the number of civil and uncivil comments as well as the unique authors involved in the discourse. Our metric approximates human judgments more accurately than previous metrics. We then use the metric to evaluate the outcomes of replies to hate speech. A linguistic analysis uncovers the differences in the language of replies that elicit follow-up conversations with high and low incivility. Experimental results show that forecasting incivility is challenging. We close with a qualitative analysis shedding light into the most common errors made by the best model.

Introduction

The pervasive problem of online hate speech has motivated researchers to investigate methods for mitigating hatred such as content moderation (Schmidt and Wiegand 2017; Fortuna and Nunes 2018). Engaging with hate speech—for example by using counterspeech—has recently emerged as an alternative to address hate speech (Richards and Calvert 2000). While content moderation consists in flagging or removing hate speech, engaging with hateful content—for example by showing sympathy towards the victim or providing counterarguments—does not interfere with the principle of free and open public spaces for debate (Mathew et al. 2019; Schieb and Preuss 2016; Chung et al. 2019).

Recently, the NLP community has focused on analyzing, modeling, and generating replies to hate speech (Mathew et al. 2019; Tekiroğlu, Chung, and Guerini 2020; Fanton et al. 2021; Yu, Blanco, and Hong 2022). These previous efforts make the following assumption: responding to hate speech is an ideal solution to stop or at least mitigate online incivility. While intuitive, we are not aware of strong evidence supporting this assumption. In this work, incivility is defined as offensive interaction ranging from abusive language, vulgarity, racism, or sexism, to harassment, name-calling or personal attacks (Antoci et al. 2016; Sadeque et al. 2019; Davidson, Sun, and Wojcieszak 2020). Consider the Reddit conversation in Figure 1.[1] The first post is hateful towards feminists in general. The second post (i.e., the reply) appears to be a strong reply that counters the hateful content. As strong as it might be, however, the second post elicits additional incivility: the third post escalates the hatred further by attacking the author of the the second post.

Hateful post: *Just curious how you can identify with a movement which has essentially become a hate group full of crazy feminists.*

Reply to hateful post: *Come on man, most feminists are ok. Hate group? how can you use such a strong term?*

Uncivil post (after the reply): *No, it's not strong. Don't lie through your teeth, c**t.*

Figure 1: An excerpt from a Reddit conversation. The second post contains counterspeech but it elicits additional uncivil behaviors. Indeed, the third post escalates the hate with respect to the original hateful post.

At face value, coming up with elaborate counterspeech replies does address online hatred. Some replies, however, may elicit additional incivility in the subsequent conversations. Existing work lacks a deeper understanding of what replies to hate speech can stop the spread of hatred and prevent uncivil content from emerging in subsequent conversations. In this paper, we aim to computationally assess and forecast conversational outcomes (namely, *conversation incivility*) of replies to hate speech. We look at all replies with varying conversational outcomes. Regardless of the content of replies—short or long, offensive or polite, well-argued or fatally flawed from a logical standpoint—we consider the outcome is civil if the discourse that follows is primarily not uncivil. Further, we argue that looking at genuine online discourse and assessing what comments elicit additional uncivil behaviors—even if they are well-meaning and polished arguments—is a worthwhile goal.

[*]Work performed while at the University of North Texas.

[1]The examples in this paper contain hateful content. We cannot avoid it due to the nature of our work.

We focus on incivility as the conversational outcome and investigate how language usage is tied to the future trajectory of a conversation. Recent studies on measuring conversation incivility use number or ratio of uncivil comments (Liu et al. 2018; Chang, Cheng, and Danescu-Niculescu-Mizil 2020; Dahiya et al. 2021; Garland et al. 2022). In contrast, we quantify conversation incivility by also considering conversation length (Tsagkias, Weerkamp, and De Rijke 2009; Yano and Smith 2010; Artzi, Pantel, and Gamon 2012) and user re-entry behaviors (Backstrom et al. 2013). This allows us to build bridges between works on conversation modeling in general and in incivility domains. Additionally, the work presented here could complement current studies on counterspeech by investigating a richer source of replies to hate speech and providing insights into strategies to respond to hatred.

We propose a new metric to assess conversation incivility and apply it to evaluate conversations following replies to hate speech in a new dataset of Reddit conversations.[2] We takes into account the number of uncivil and civil comments in a conversation and also considers user behavior. Our metric differentiates replies depending on the outcomes of the follow-up conversation, for example, comments by many different "drive-by" people and those by relatively few people who engage multiple times (Backstrom et al. 2013). Intuitively, a reply to hateful content is more divisive and problematic if it attracts more hate speakers to participate, misbehave, or name-call. Additionally, as we shall see, the metric approximates human judgments more accurately than previous metrics. For instance, the conversation following a reply is less civil if it consists of five uncivil posts instead of one, despite the ratio of uncivil posts is the same. Similarly, if the five uncivil posts are from different users, the conversation is less civil than if they are from the same user, despite the number and ratio of uncivil posts are the same.

We focus on conversation incivility following replies to hate speech in Reddit. A linguistic analysis is presented to show the differences in the language of replies to hate speech depending on the incivility of the follow-up conversation. Importantly, we show that the linguistic insights hold across several Reddit communities, including some that are known for respectful debate rather than hateful content (e.g., *r/ChangeMyView*). We then experiment with classifiers to predict whether a reply will be followed by a conversation with high, medium or low incivility. Our models obtain modest results, and we present a qualitative error analysis.

In summary, we answer the following research questions:

1. Do all replies to hateful posts elicit conversations with the same incivility? (they don't);
2. Do replies that elicit high, medium, and low conversation incivility use different language? (they do);
3. Do models to predict the incivility of the conversation following replies to hateful post benefit from having access to the hateful post in addition to the reply? (they do);
4. When differentiating between replies eliciting the top-k highest and lowest incivility, is it true that the smaller the k the easier the task? (it is).

[2] Data available at https://github.com/xinchenyu/incivility

Related Work

Replying to Hate Speech There is an abundance of research working on replies to hate speech. Most prior work focuses on counterspeech and contribute several corpora with counterspeech content (Mathew et al. 2019; Qian et al. 2019; Chung et al. 2019; Yu, Blanco, and Hong 2022). Chung et al. (2019) collect synthetic counterspeech generated on-demand by trained operators. Compared to genuine counterspeech written by regular people out of their own desires and motivations, synthetic counterspeech is more generic and may not address specific hateful content (e.g., *This kind of language is inappropriate and should be avoided*). Some user studies that investigate the outcome of replies to hatred lack in scale; they compare the control group with the treatment group that has received interventions (Munger 2017; Hangartner et al. 2021; Bilewicz et al. 2021; Wachs et al. 2023). The only two large-scale previous works are by Garland et al. (2022) and Albanyan, Hassan, and Blanco (2023). Garland et al. (2022) calculate the hate score of each reply and estimate its impact by comparing the average hate scores before and after. Albanyan, Hassan, and Blanco (2023) focus on the direct replies after each reply to hatred. While we are inspired by previous approaches addressing online hatred, we analyze *all* user-generated comments after replies to hate speech *in genuine online conversations*. Additionally, measuring incivility automatically allows us to bypass the burden of manual annotations and work with orders of magnitude more data than previous work.

Conversational Forecasting There are several efforts on forecasting whether online content will result in additional uncivil behaviors. Cheng et al. (2017) predict whether a moderator will flag a post for removal. Zhang et al. (2018) and Yuan and Singh (2023) predict whether a few utterances at the very beginning of a conversation will lead to a personal attack. Liu et al. (2018) forecast whether an Instagram post will receive more than n uncivil comments. Dahiya et al. (2021) forecast the incivility score of upcoming tweet replies. Similarly, our work aims to forecast incivility of the conversation following a reply to hate speech. We, however, measure incivility by also considering the total number of comments as well as user re-entry behaviors in a conversation. As we shall see, our metric to measure conversation incivility better approximates human judgments.

Besides conversation incivility, a rich source of other conversational outcomes have been explored, including betrayal in games (Niculae et al. 2015), success in persuading others (Tan et al. 2016), debate winners (Potash and Rumshisky 2017), online conflicts (Levy et al. 2022), and prosocial behaviors (Bao et al. 2021; Lambert, Rajagopal, and Chandrasekharan 2022). We build on prior work in modeling conversation trajectory from the structural aspects of conversations (Backstrom et al. 2013) and complement them by considering linguistic aspects.

Measuring Conversation Incivility

We propose a new metric to measure conversation incivility following a reply r. Our metric consists of two main components: uncivil behavior $U(r)$ and civil behavior $C(r)$. A comment is either civil or uncivil. All comments in the subsequent conversation after a reply r are from a population of unique users P. For each user i in P (for $i = 0, 1, 2, ..., k$), let n_{ui} denote the number of uncivil comments the user P_i posts, and n_{ci} denote the number of civil comments. Uncivil behavior $U(r)$ is defined as the sum of $f(n_{ui})$ over all the users i in the conversation after r. Similarly, civil behavior $C(r)$ is defined as the sum $f(n_{ci})$ over all the users. Here, f is a strictly increasing and concave down function passing through the origin. When there are no comments after a reply r, the conversation incivility is equal to 0. Intuitively, the conversation incivility following a reply r is higher when there are (a) more uncivil comments by many people and (b) fewer civil comments by a handful of people.

We formally define the conversation incivility score of r ($S(r)$) as follows:

$$S(r) = \alpha U(r) - (1 - \alpha)C(r)$$

where $U(r) = \sum_{i=1}^{k} f(n_{ui})$ and $C(r) = \sum_{i=1}^{k} f(n_{ci})$.

The parameter α determines how much weight to give to each component. The larger α is, the more weight is given to uncivil behavior (i.e., more civil comments are needed to neutralize one uncivil comment). Following the literature, the future trajectory of a conversation becomes more toxic and unhealthy if it involves a larger number of participants venting or misbehaving, and this should be treated differently from situations in which there are only few people with repeated engagement (Backstrom et al. 2013). Finally, a reply eliciting a long and civil conversation reflects that it attracts a great deal of attention and promotes healthy discussions. Our metric builds on prior work on comment-volume prediction (Artzi, Pantel, and Gamon 2012; Backstrom et al. 2013), but we adapt it to model conversation incivility.

In this paper, we experiment with the square root function as f (i.e., $f(x) = \sqrt{x}$), but any strictly increasing and concave down function is a valid choice. The rate of increase of these functions slows as x grows; different f choices dictate how quickly additional comments by the same author are deemed unimportant. As shown in the appendices, the Spearman's correlation coefficients between $S(r)$ using $\sqrt{x}$ and other choices ($\log(x)$, $\sqrt[3]{x}$, $arctan(x)$, $tanh(x)$) are over 0.97 ($p < 0.001$). In other words, the choice of f is not critical. Indeed, while the absolute scores of conversation incivility will vary depending on f, comparing scores would lead to the same conclusions.

A Corpus of Online Hate, Replies, and Follow-Up Conversation Incivility

We choose Reddit as the starting point for our corpus and use the PushShift API to retrieve whole conversation threads.[3] As the prevalence of online hate in the wild is very low (0.1% in English language social media (Vidgen et al. 2019)), many studies use keyword sampling to increase the chances. Keywords, however, may introduce topic and author biases (Wiegand, Ruppenhofer, and Kleinbauer 2019; Vidgen et al. 2021). Instead, we use community-based sampling and identify 39 subreddits that are thought to be more or less hateful (Qian et al. 2019; Guest et al. 2021; Vidgen et al. 2021) or previously used in conversational forecasting (Chang and Danescu-Niculescu-Mizil 2019; Yuan and Singh 2023) (see Appendices for the full list). Note that some subreddits we work with primarily consist of respectful conversations rather than hateful content (e.g., *r/technology*, *r/changemyview*). We retrieve 1,382,596 comments from 5,325 submissions in the 39 subreddits. The publication time of the Reddit conversations range from October 16, 2020, to February 20, 2022.

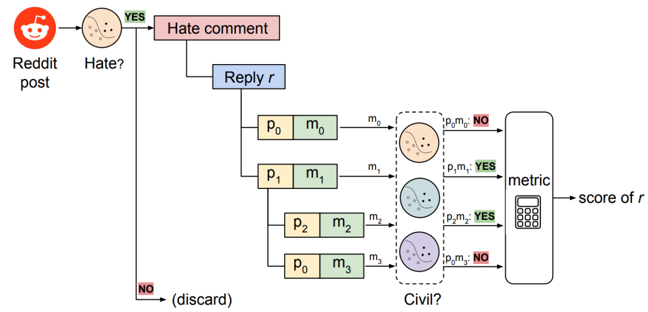


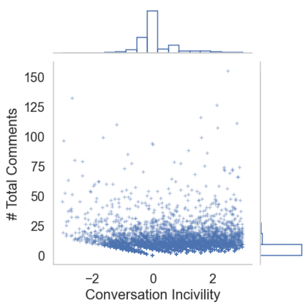
Figure 2: The pipeline to create our corpus. We show a Reddit post, the reply to it, and the subsequent conversation. The output is the incivility score of the conversation following the reply. We indicate each comment in the subsequent conversation as m_j (for $j = 0, 1, .., h$) and the user who posts it as p_i (for $i = 0, 1, .., k$). The pipeline includes two questions: (1) is the Reddit post hateful? and (2) is m_j civil? We adopt pretrained classifiers to answer the questions.

We focus on replies that directly reply to hate speech. Therefore, the next steps are to (a) identify the comments that are hateful and their replies, and (b) assess the incivility of the follow-up conversations. The second step requires identifying uncivil content in the comments following the reply. Figure 2 illustrates the process.

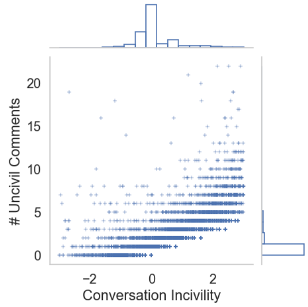
Identifying Hate Comments and Their Replies We identify hate speech in the 1,382,596 comments using pretrained models (Liu et al. 2019) fine-tuned with the corpus by Qian et al. (2019) and the implementation by Pruksachatkun et al. (2020). We make this choice for several reasons. First, the corpus annotates Reddit comments as hateful or not hateful, the same domain we work with. Second, the classifier obtains outstanding results: 0.93 F1. In a more strict evaluation using Cohen's κ, we obtain $\kappa = 0.83$ between the predictions and the gold annotations in the test set. Note that κ coefficients above 0.80 indicate (almost) perfect agreement (Artstein and Poesio 2008). Thus the predictions are reliable enough to be considered ground truth.

After automatically identifying hateful comments, we pair (a) each hateful comment with each of its direct replies and (b) each reply to a hateful comment with the follow-up conversation (i.e., all the subsequent comments in the same thread). We found 21,845 hate comments in the 39 subreddits we work with. On average, a hate comment has 1.56

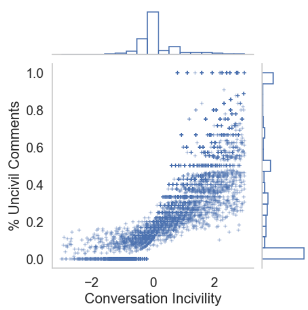
[3]https://pushshift.io/api-parameters/



(a) # Total comments



(b) # Uncivil comments



(c) % Uncivil comments

Figure 3: Comparison of our conversation incivility metric and three prior metrics: (a) number of total comments, (b) number of uncivil comments, and (c) percentage of uncivil comments. Our metric distinguishes between many conversations that obtain the same scores with prior metrics.

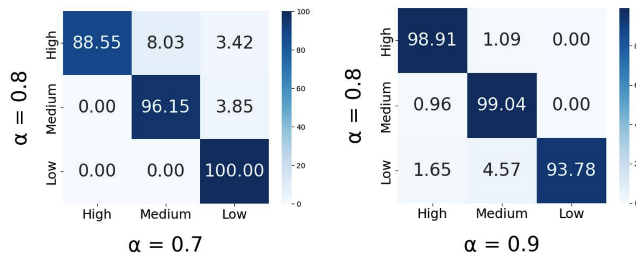


Figure 4: Confusion matrix (percentages) showing label changes (High, Medium, or Low incivility) when α is 0.8 vs. 0.7 and 0.9 respectively.

Hate: *Lol this thread is full of internet losers like you.* Reply: *Ha! You don't even know me little man.*	
High incivility	$S(r) = 5.02$, [u:26, c:7]
Hate: *Too easy to trigger you maga f**ks, snowflake.* Reply: *Have a nice day!*	
Medium incivility	$S(r) = 0$, [u:0, c:0]
Hate: *Trash talking is perfectly ok. He is an a**hole.* Reply: *It's *tolerable* under those conditions. It is not perfectly okay.*	
Low incivility	$S(r) = -3.92$, [u:3, c:29]

Table 1: Examples from our corpus and their incivility levels. We also include their incivility scores and number of uncivil (u) and civil comments (c).

direct replies, and there are 2.36 comments in the conversation following up a reply to hateful content.

Assessing Conversation Incivility The metric to measure conversation incivility following a reply r requires us to get the number of uncivil and civil comments published after r by each user. Compared with hateful comments, uncivil comments include broader cases (Davidson, Sun, and Wojcieszak 2020). We therefore calculate uncivil comments based on the output of three classifiers. We build three models by training the same architecture as before with the corpus by Qian et al. (2019) and two additional corpora (Davidson et al. 2017; Vidgen et al. 2021). We consider a comment published after r as uncivil if any of the three classifiers predicts *uncivil*. Otherwise, we consider it *civil*. After calculating uncivil and civil behavior after each reply, calculating the conversation incivility score $S(r)$ is straightforward. We experiment with $\alpha = 0.8$, as we consider one uncivil content more critical and several civil messages are needed to neutralize uncivil content. We do not claim that this is the best choice. Instead, we argue that α ought to be chosen based on the level of uncivil behaviors that is acceptable in a conversation, and can be adjusted according to the social media platform or tolerance to incivility. As we shall see, $\alpha = 0.8$ leads to almost perfect agreement with human annotators.

The scatter plots in Figure 3 compare our conversation incivility metric with three prior metrics: (a) number of (total) comments, (b) number of uncivil comments, and (c) percentage of uncivil comments. The three prior metrics often assign the same incivility scores to many conversations. For example, the vast majority of conversations following a reply to hate speech are very short (few total comments, Figure 3a) and do not have any uncivil comments (Figure 3b). Similarly, the percentage of uncivil comments is low, although this metric is less biased towards low scores. Our conversation incivility metric (x-axis in the three plots) assigns different scores to conversations that receive the same scores with the three prior metrics, thereby providing a more nuance distinction of civil and uncivil conversations.

Manual Validation To validate the conversation incivility scores obtained by our metric, we manually annotate a small benchmark. Specifically, we create a benchmark with ground-truth annotations in four steps. First, we randomly select 500 replies to hate comments from our corpus such that the follow-up conversation has at least one comment. Second, we retrieve Reddit snippets containing the hateful comment, the reply, and the follow-up conversation. Third, we randomly pair the 500 snippets, resulting in 250 pairs. Fourth, annotators manually annotate which of the follow-up conversations in each pair is more uncivil. Two research assistants were hired as annotators. 10 pairs were discarded due to uncertainty; annotators agreed on 194 of the remaining 240 pairs (80.8%). The Cohen's κ coefficient is 0.62, which is considered *substantial* agreement (Artstein and

	All	Discussion	Hobby	Identity	Meme	Media
Textual factors						
Tokens	↑↑↑	↑↑↑	$\underline{\uparrow\uparrow}$	↑↑↑	↑↑↑	↑↑↑
Negations	↑↑↑	↑↑↑	$\underline{\uparrow}$	↑↑↑	↑↑↑	$\underline{\uparrow\uparrow\uparrow}$
1st person pronouns	↑↑↑	$\underline{\uparrow\uparrow\uparrow}$	$\underline{\uparrow}$	↑↑↑	↑↑↑	$\underline{\uparrow\uparrow}$
2nd person pronouns	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑
Named entity (norp)	↑↑↑	↑↑↑	$\underline{\uparrow}$	$\underline{\uparrow}$		$\underline{\uparrow}$
Question marks	↑↑↑	$\underline{\uparrow}$		$\underline{\uparrow\uparrow\uparrow}$	$\underline{\uparrow}$	$\underline{\uparrow}$
Quotations	↑↑↑	↑↑↑	$\underline{\uparrow}$	$\underline{\uparrow\uparrow}$		$\underline{\uparrow}$
Sentiment factors						
Positive words	↓↓↓	↓↓↓	$\underline{\downarrow\downarrow\downarrow}$	↓↓↓	$\underline{\downarrow\downarrow}$	$\underline{\downarrow\downarrow}$
Negative words	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑
Disgust words	↑↑↑	$\underline{\uparrow\uparrow\uparrow}$	$\underline{\uparrow\uparrow\uparrow}$	↑↑↑	$\underline{\uparrow}$	$\underline{\uparrow}$
Hatred words	↑↑↑	$\underline{\uparrow}$		$\underline{\uparrow\uparrow}$		$\underline{\uparrow\uparrow\uparrow}$
Angry words	↑↑↑	$\underline{\uparrow\uparrow}$	$\underline{\uparrow}$	$\underline{\uparrow\uparrow}$	$\underline{\uparrow}$	$\underline{\uparrow\uparrow}$

Table 2: Linguistic analysis comparing the replies to hateful comments that have high and low conversation incivility. We analyze all Reddit conversations (Column 2) and each subreddit category (Column 3-7, see the list in the Appendices). Number of arrows indicates the p-value (t-test; one: $p < 0.05$, two: $p < 0.01$, and three: $p < 0.001$). Arrow direction indicates whether higher values correlate with high (up) or low (down) incivility. Tests that *do not pass* the Bonferroni correction are underlined.

Poesio 2008). We include in our benchmark the 194 pairs with perfect agreement.

Armed with the benchmark, we compare the ground truth (i.e., which of the two follow-up conversations in each pair is more uncivil according to the human annotators) with (a) our conversation incivility metric and (b) the best prior metric (i.e., percentage of uncivil comments, Figure 3). Using the scores obtained with our metric (higher score indicates more uncivil), we match the ground truth in 183 out of the 194 pairs (94.3%), resulting in a Cohen's κ of 0.89, which is almost perfect agreement. On the other hand, using the percentage of uncivil comments, we match the ground truth in 172 out of the 194 pairs (88.7%), resulting in a Cohen's κ of 0.77, which is substantial agreement. Based on this evaluation, we conclude that while the percentage of uncivil comments is a valid choice to approximate conversation incivility, our conversation incivility metric more closely approximates humans judgments thus it is more sound.

Conversation Incivility Level We use the scores to group all replies in our corpus based on quantile (top 25%, middle 50%, and bottom 25%). The score ranges are as follows:

- *High* incivility with $S(r) \in (0.22, 16.26]$;
- *Medium* incivility with $S(r) \in (-0.20, 0.22]$; and
- *Low* incivility with $S(r) \in [-48.76, -0.20]$.

We refer to this grouping (high, medium or low) as the conversation incivility level. Figure 4 shows the changes in incivility levels depending on other choices of α (0.7 and 0.9). Most conversations are assigned the same incivility level. We note that absolute incivility scores are not critical. Rather, incivility scores should be used as a means to compare the incivility of two conversations. The following analyses are based on the conversation incivility levels.

Corpus Analysis

Our final corpus consists of 34,115 replies to hateful comments from Reddit along with the conversation following each reply. We show examples of each incivility level in Table 1. In the first example, the reply shows disagreement by attacking the author of the hate comment (*little man*). This counter hate reply leads to high conversation incivility: 26 out of 33 comments that follow the reply are uncivil. Indeed, we observe the follow-up conversation is made up of behaviors where the original authors of the hateful comment and the reply repetitively denigrate each other. In the second example, the reply uses sarcasm to stop arguing with the hateful comment ("*Have a nice day!*"). There are no comments after this reply, yielding a 0 incivility score. It is followed by *medium* conversation incivility—no additional comment is posted. In the third example, the reply denounces the misbehavior in the hate comment is inappropriate without getting into details or personal attacks ("*[...] It is not perfectly okay.*"). This is a common counter hate strategy (Mathew et al. 2019) and the follow-up conversation is barely uncivil: although 3 comments are uncivil, they are posted by the same user and 29 comments from 24 users remain civil.

Linguistic Insights We perform a linguistic analysis to shed light on differences in the language people use in the replies that elicit high and low conversation incivility (Table 2). As different Reddit communities vary in content due to differences in topics, moderation rules, etc. (Weld, Zhang, and Althoff 2021, 2022), we further conduct analyses for each type of community to explore whether the differences between high and low incivility across communities exist. The 39 subreddits are grouped into five categories based on the taxonomy from Weld, Zhang, and Althoff (2022). The authors hand-labeled communities in the categories iteratively until reaching agreement among them. We include News in Media-sharing as both communities are for sharing information. The five categories are: Discussion (*e.g., r/antiwork*), Hobby (*e.g., r/dota2*), Identity (*e.g., r/Feminism*), Meme (*e.g., r/DankMemes*), and Media-sharing (*e.g., r/worldnews*) (see Appendices).

All factors we consider are based on counts of (a) textual features (top block) or (b) presence of words related to sentiment. We consider a reply uses quotations if it has the character '>' and the text that follows overlaps with the hateful comment (Chakrabarty et al. 2019; Jo et al. 2020). We check for negation cues using the list by Fancellu, Lopez, and Webber (2016). We use spaCy to recognize named entities.[4] For sentiment and cognition, we use the Sentiment Analysis and Cognition Engine (SEANCE) lexicon, a tool for psychological linguistic analysis (Crossley, Kyle, and McNamara 2017). Statistical tests are conducted using unpaired t-tests between two groups: the replies eliciting high and low incivility. We draw several interesting insights:

- Regarding textual factors, we observe that among all the replies, the more tokens, negations, pronouns (1st and 2nd person), entities (nationalities, religious, or political groups), question marks and quotations, the more likely the subsequent conversation of a reply to hateful comment is uncivil. Negation cues are often used to dispute the hateful comment. Presence of *you* and *your* usually refers to the author of the hateful comment.
- Regarding sentiment, there are significant differences in the uses of positive and negative words. Replies that use more hatred, disgust and angry words tend to lead to more incivility in the follow-up conversations.
- Although topics and content vary across communities, we observe high consistency in the differences between high and low conversation incivility, especially for 2nd person pronouns and negative words.

Experiments and Results

We experiment with models to solve two problems:

- Determining the conversation incivility level of a reply to hate speech: high, medium or low incivility; and
- Differentiating the top-$k\%$ and bottom-$k\%$ replies according to their conversation incivility scores.

All our models are neural classifiers with the RoBERTa transformer (Liu et al. 2019) as the main component. We use the pretrained models by HuggingFace (Wolf et al. 2020) and Pytorch (Paszke et al. 2019) to implement our models.

Determining Incivility Level

The neural architecture consists of the RoBERTa transformer, a fully connected layer (768 neurons and `tanh` activation), and another fully connected layer (3 neurons and softmax activation) to make predictions (high, medium, or low incivility). To investigate whether adding the hate comment would be beneficial, we consider three textual inputs:

- the hate comment;
- the reply to the hate comment; and
- the hate comment and the reply.

Intuitively, the reply is the most important input, but as we shall see including the hate comment is beneficial. We concatenate both inputs with the `[SEP]` special token.

Pretraining with Related Tasks We experiment with several corpora to investigate whether pretraining with related tasks is beneficial. Specifically, we pretrain with existing corpora annotating: (a) hate speech: hateful or not hateful (Davidson et al. 2017); (b) sentiment: negative, neutral, or positive (Rosenthal, Farra, and Nakov 2017); (c) sarcasm: sarcasm or not sarcasm (Ghosh, Vajpayee, and Muresan 2020); (d) counterspeech: hate, neutral, or counterhate (Yu, Blanco, and Hong 2022); and (e) stance: agree, neutral, or attack (Pougué-Biyong et al. 2021).

Blending Additional Annotations Pretraining takes place prior to training with our corpus. We also experiment with a complementary approach: blending additional corpora during the training process, as proposed by Shnarch et al. (2018). With blending, there are two phases in the training process: (a) m blending epochs using all of our corpus and a fraction of an additional corpus, and (b) n epochs using only our corpus. In each blending epoch, a random fraction of an additional corpus is fed to the network. The fraction is determined by a blending factor $\alpha \in [0..1]$. The first blending epoch is trained with our corpus and the whole additional corpus. Subsequent blending epochs use smaller fractions of the additional corpus. We use for blending purposes the corpora we use for pretraining that annotate three labels (Rosenthal, Farra, and Nakov 2017; Pougué-Biyong et al. 2021; Yu, Blanco, and Hong 2022).

Quantitative Results We split the 34,115 replies in our corpus into train (60%), development (20%) and test (20%) splits. We present results with the test split in Table 3. The majority baseline always predicts *medium*. The remaining rows present results with different settings: using as input the *hate comment*, the *reply* or both without pretraining or blending, and also with pretraining, blending and both. We provide here results pretraining and blending with the most beneficial tasks: counterspeech (+ *pretraining* when using *reply*, and + *blending* using *hate comment* + *reply*), and stance for the remaining ones. We tune the blending factor α with the training and development splits, like other hyperparameters. We found the optimal α to be 0.5 when only blending and 1.0 when pretraining and blending.

Using only the reply as input is a strong baseline: it substantially outperforms the majority baseline (F1: 0.44 vs. 0.32). Using both the hate comment and reply yields better results (F1: 0.46). Pretraining and blending yield better results compared with blending alone (*reply*: 0.49 vs. 0.45, *hate comment* + *reply*: 0.52 vs. 0.46). Also, pretraining or pretraining and blending are more beneficial when the input is both the hate comment and the reply. Finally, the networks (a) pretraining and (b) pretraining and blending using both hate comment and reply yield the best results (F1: 0.52).

Differentiating between the Top-$k\%$ and Bottom-$k\%$ replies

Although determining the incivility level of any replies to hateful comment is a worthwhile goal, differentiating between the top-$k\%$ and bottom-$k\%$ replies according to the incivility scores of their follow-up conversations may lead to better actionable knowledge. Indeed, replies eliciting the highest (top-$k\%$) or lowest (bottom-$k\%$) incivility scores are more informative than the large amount of replies that elicit conversations with in-between incivility scores.

[4] https://spacy.io/usage/linguistic-features

	High			Medium			Low			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority Baseline	0.00	0.00	0.00	0.49	1.00	0.66	0.00	0.00	0.00	0.24	0.49	0.32
RoBERTa classifier with												
hate comment	0.34	0.31	0.33	0.53	0.72	0.61	0.27	0.10	0.15	0.41	0.46	0.42
reply	0.42	0.33	0.37	0.53	0.77	0.63	0.29	0.10	0.15	0.44	0.49	0.44
+ blending	0.42	0.42	0.42	0.55	0.76	0.63	0.33	0.08	0.13	0.46	0.50	0.45
+ pretraining†	0.44	0.37	0.40	0.56	0.72	0.63	0.32	0.20	0.25	0.47	0.50	0.47
+ blending†‡	0.47	0.39	0.43	0.57	0.70	0.63	0.34	0.25	0.29	0.49	0.51	0.49
hate comment + reply	0.43	0.32	0.36	0.55	0.66	0.60	0.32	0.27	0.29	0.46	0.48	0.46
+ blending†	0.43	0.37	0.40	0.54	0.79	0.64	0.35	0.10	0.16	0.47	0.51	0.46
+ pretraining†‡	**0.52**	**0.43**	**0.47**	**0.59**	**0.77**	**0.67**	0.42	0.23	0.30	**0.53**	**0.55**	**0.52**
+ blending†‡	0.48	0.42	0.45	0.59	0.72	0.65	**0.38**	**0.27**	**0.32**	**0.51**	**0.53**	**0.52**

Table 3: Results obtained with several models. We indicate statistical significance (McNemar's test (McNemar 1947) over the weighted average) as follows: † indicates statistically significant ($p < 0.05$) results with respect to the *reply* model, and ‡ with respect to the *hate comment* + *reply* model. Training with the *hate comment* + *reply* coupled with pretraining with stance or both pretraining and blending stance yields the best results (F1: 0.52).

		Top-k%			Bottom-k%			Weighted Average		
	Size	P	R	F1	P	R	F1	P	R	F1
$k = 5$	3,234	0.77	0.74	0.76	0.72	0.75	0.74	0.75	0.75	0.75
$k = 10$	6,124	0.74	0.77	0.76	0.68	0.65	0.67	0.72	0.72	0.72
$k = 15$	10,434	0.70	0.76	0.73	0.68	0.61	0.65	0.69	0.69	0.69
$k = 20$	17,231	0.67	0.63	0.65	0.65	0.68	0.66	0.66	0.66	0.66

Table 4: Experimental results differentiating the top-$k\%$ and bottom-$k\%$ replies to hateful comments according to the incivility scores in the follow-up conversations. We present results for several values of k. The results are higher than when also identifying replies that elicit *medium* incivility. Additionally, it is easier to differentiate the replies with the highest and lowest incivility scores in the follow-up conversations: the smaller the k, the higher the weighted average.

Table 4 presents the results with several k values after retraining one of the best performing systems from Table 3 (*hate comment* + *reply* + *pretraining*). Results show that the smaller the k, the easier it is to differentiate between the two kinds of replies. This is especially true when $k = 5$ (F1: 0.75). The results are encouraging. Indeed, replies eliciting the highest and lowest conversation incivility differ in language usage and the classifier can distinguish them. Also, these replies are the ones with the potential to elicit the highest and lowest incivility in the subsequent conversation, thus are the most useful to identify.

Qualitative Analysis

When determining the incivility level of a reply, when does our best model (Table 3) make mistakes? To investigate this question, we manually analyze 200 random samples in which the output of the network differs from the ground truth. Table 5 exemplifies the most common error types.

The most frequent error type (23%) is *Rhetorical questions*, a finding consistent with previous work (Schmidt and Wiegand 2017). In the example, the model fails to realize that the question in the reply is used to point out inappropriate content rather than expecting an answer. There are no comments after the reply, therefore the conversation that follows has medium incivility.

The second and third most common error types (18% and 16%) occupied when a reply is (a) uncivil but followed by a conversation with low incivility or (b) civil but followed by a conversation with high incivility. Using uncivil language is a counter hate strategy (Mathew et al. 2019), and the model fails to recognize when doing so leads to low conversation incivility. Similarly, the model struggles when countering hate politely elicits additional uncivil behaviors. When correcting misstatements, language toxicity may increase (Mosleh et al. 2021).

Sarcasm and irony are also common error types (15%) in our task. This finding is consistent with previous work on detecting hate (Nobata et al. 2016; Qian et al. 2019). In the example, using sarcasm to point out a bad argument (i.e., name calling) elicits further hate and the model errs.

Errors may also occur (10%) when general knowledge is required to identify hate content that does not use offensive language (e.g., calling people *incels*).

Finally, we observe that *negation* appears in 8% errors. In the example, negations are used to point out the flaws of generalizing. We hypothesize that the model fails to identify that the reply is followed by medium incivility: negation does indicate high incivility in general (Table 2).

Error Type	%	Example	Ground Truth	Predicted
Rhetorical question	23	Hate: *You're living in the west. You're privileged.* Reply: *Are you an idiot? Can you read? Feminist my a**.*	Medium	High
Uncivil reply followed by conversation with low incivility	18	Hate: *I've addressed this about forty times with as many smooth brains as you so. You're an idiot.* Reply: *I think you might be the idiot here retard.*	Low	High
Civil reply followed by conversation with high incivility	16	Hate: *You're an ignorant twat who just parrots what they read in FB and reddit memes. [...] What a cancer you are.* Reply: *Calling others cancer is taking it too far. Mind rule 4, please.*	High	Low
Sarcasm or irony	15	Hate: *No you retard, where is the f**king lie?* Reply: *Name calling nice argument.*	High	Low
General knowledge	10	Hate: *lol bet you thought a single thing you said wasn't retarded.* Reply: *This place is infested with incels and TD trolls.*	High	Medium
Negation	8	Hate: *Why we have to tolerate Islam? They call us filth. Christians are horrible as well. Both are f**king awful.* Reply: *Not all Muslims are bigots, like not all Christians are bigots.*	Medium	High

Table 5: Most common error types made by the best model (predictions by *hate comment* + *reply* + *pretraining*).

Conclusion and Discussion

In this work, we present a metric to assess conversation incivility and apply it to conversations following replies to hate posts in a large Reddit dataset. Our metric takes into account the number of both civil and uncivil comments as well as the unique authors. A manual validation shows that out metric approximates human judgments better than previous proposals. Regardless of whether replies convincingly counter a hate post, we believe it is worthwhile to identify what kind of user-generated content attracts attention and shapes civil discussions. While we make no causal claims about which linguistic features could affect conversation incivility, we show that the language of user-generated replies differs depending on their conversation incivility levels. The outcomes of the linguistic analysis are intuitive. For example, replies that use more negative and disgust words result in follow-up conversations with higher incivility. This insight is consistent across several communities. Experimental results show that pretraining with and blending existing corpora yield improvements, yet the task of forecasting conversation incivility is still challenging to automate.

Experimental results with classifiers built to distinguish the replies that elicit follow-up conversation with the top-$k\%$ and bottom-$k\%$ incivility scores are encouraging. The smaller the k, the easier the task, despite the classifiers do not have access to the follow-up conversation. The work presented here opens the door to automated methods to forecast the incivility of the conversation following a reply to hate content—*at the time* the reply is posted. It also may inform the design of effective strategies to mitigate the spread of hate *without* having to censor hateful content.

Broader Perspectives, Ethics, and Competitive Interests

Our work provides both theoretical and empirical guidelines to assist online media in measuring, understanding, identifying, and even intervening with content that could elicit additional incivility. First, our research proposes a more comprehensive approach to measure conversational outcomes with regard to incivility, compared with traditional approaches using only the number or ratio of uncivil comments. Second, our study provides language characteristics that could trigger additional incivility, revealing deeper insights into the nature of these conversations.

This work may motivate the design of systems that highlight content likely to elicit uncivil behaviors. Doing so could assist moderation. For instance, content that does not contain swear words or use implicit hate speech (ElSherief et al. 2021) have instigated additional uncivil behaviors and violated Reddit content policy.[5] This poses a challenge for most detection systems. However, our method could help identify this kind of content. In addition, our metric to estimate conversation incivility offers a new way to measure conversational outcomes, focused on the future health of the conversation. Our metric may be applied to broader topics (e.g., persuasion, online debate) and other online discussion forums. Finally, our work may also help design guidelines for how to appropriately engage in online discussions (e.g., pointing out what kinds of strategies are the most effective when countering online hatred (Mathew et al. 2019)).

Limitations This work is not without limitations. First, our linguistic analyses should not be interpreted as causal statements. Small-scale user studies could provide an understanding of how humans perceive hate speech and replies to hate speech. Additionally, some of the linguistic analyses relies on automated tools (e.g., spaCy) that do not always output the correct predictions (e.g., some NORP named entities are false positives). Second, we identify uncivil comments automatically with classifiers. These classifiers obtain good results but are not perfect. Third, we focus on forecasting conversation incivility only from language (the hate

[5]https://www.redditinc.com/policies/content-policy

content and the reply). A promising line of future research could consider combining structural and linguistic features and incorporate other factors such as roles and user traits. Finally, we only consider the hateful comment and the reply in our experiments. More complex modeling that takes into account additional context (e.g., the full conversation prior to the reply) may be beneficial.

Ethical Considerations The study has been through careful consideration of the risks and benefits to ensure that it is conducted in an ethical manner. First, we use the PushShift API to collect data from Reddit.[6] The collection process is consistent with Reddit's Terms of Service. Second, in contrast to private messaging services, Reddit is considered a public space for discussion (Vidgen et al. 2021). It does not require IRB review. Users consent to have their data made available to third parties including academics when they sign up. Ethical guidelines state that in this situation explicit consent is not required from each user (Procter et al. 2019). We obfuscate user names to reduce the possibility of identifying users. In compliance with Reddit's policy, we would like to make sure that our dataset will be reused for non-commercial research only.[7] Third, the annotators were warned of the potential hateful content before working on our task. They were also encouraged to stop the annotation process whenever feel upset. We provide annotators with access to supporting services throughout the task. Annotators were compensated with $8 per hour. Finally, we acknowledge the risk associated with releasing the dataset. However, we believe the benefit of shedding light on what replies elicit additional incivility outweighs any risks associated with the dataset release.

Subreddit List

We provide here the list of subreddits we work with and the categories they belong to:

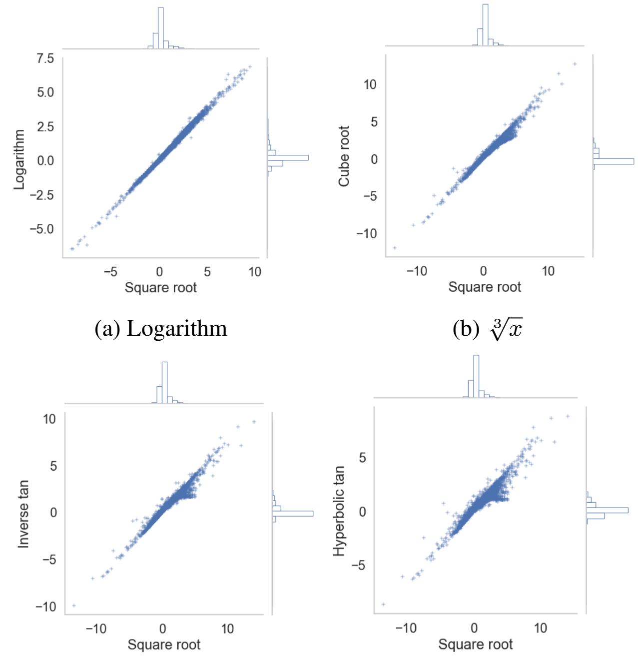
- Discussion: *r/antiwork*, *r/changemyview*, *r/NoFap*, *r/Seduction*, *r/PurplePillDebate*, *r/ShitPoliticsSays*, *r/PurplePillDebate*, *r/bindingofisaac*, *r/FemaleDatingStrategy*, *r/SubredditDrama*;
- Hobby: *r/KotakuInAction*, *r/DotA2*, *r/technology*, *r/modernwarfare*, *r/playrust*, *r/oblivion*;
- Identity: *r/bakchodi*, *r/Feminism*, *r/PussyPass*, *r/MensRights*, *r/Sino*, *r/BlackPeopleTwitter*, *r/india*, *r/PussyPassDenied*, *r/TwoXChromosomes*, *r/GenZedong*, *r/antheism*;
- Meme: *r/4Chan*, *r/justneckbeardthings*, *r/HermanCainAward*, *r/MetaCanada*, *r/DankMemes*, *r/ShitRedditSays*;
- Media: *r/conspiracy*, *r/worldnews*, *r/Drama*, *r/TumblrInAction*, *r/lmGoingToHellForThis*, *r/TrueReddit*.

Choice of f Function

Our conversation incivility metric uses a strictly increasing and concave down f function passing through the origin. We work with $f(x) = \sqrt{x}$, but the choice is not critical. Figure 5 compares the incivility scores obtained with $f(x) = \sqrt{x}$ and four alternative f functions: logarithm ($\log(x+1)$), cube root ($\sqrt[3]{x}$), inverse tangent ($\arctan(x)$), and hyperbolic tangent ($\tanh(x)$). Spearman's correlation coefficients are 0.99, 0.99, 0.99, and 0.97 respectively ($p < 0.001$). Thus any function would yield the same outcomes if incivility scores, as we recommend, are to be used only for comparative purposes (i.e., in relative rather than absolute terms).

[6]https://pushshift.io/api-parameters/

[7]https://www.reddit.com/wiki/api-terms/



(a) Logarithm (b) $\sqrt[3]{x}$ (c) $\arctan x$ (d) $\tanh x$

Figure 5: Comparison of incivility scores obtained with our choice of f function ($f(x) = \sqrt{x}$) and other strictly increasing and concave down functions passing through the origin: logarithm ($\log(x+1)$), cube root ($\sqrt[3]{x}$), inverse tangent ($\arctan(x)$), and hyperbolic tangent ($\tanh(x)$). Spearman's rank correlation coefficients are greater than 0.97 ($p < 0.001$), indicating the the choice of f is not critical, especially since incivility scores are meant to be used for comparison purposes (i.e., relative terms).

References

Albanyan, A.; Hassan, A.; and Blanco, E. 2023. Not All Counterhate Tweets Elicit the Same Replies: A Fine-Grained Analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, 71–88. Toronto, Canada: Association for Computational Linguistics.

Antoci, A.; Delfino, A.; Paglieri, F.; Panebianco, F.; and Sabatini, F. 2016. Civility vs. incivility in online social interactions: An evolutionary approach. *PloS one*, 11(11): e0164286.

Artstein, R.; and Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.*, 34(4): 555–596.

Artzi, Y.; Pantel, P.; and Gamon, M. 2012. Predicting Responses to Microblog Posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 602–606. Montréal, Canada: Association for Computational Linguistics.

Backstrom, L.; Kleinberg, J.; Lee, L.; and Danescu-Niculescu-Mizil, C. 2013. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-Entry. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, 13–22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318693.

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*, WWW '21, 1134–1145. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Bilewicz, M.; Tempska, P.; Leliwa, G.; Dowgiałło, M.; Tańska, M.; Urbaniak, R.; and Wroczyński, M. 2021. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive behavior*, 47(3): 260–266.

Chakrabarty, T.; Hidey, C.; Muresan, S.; McKeown, K.; and Hwang, A. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2933–2943. Hong Kong, China: Association for Computational Linguistics.

Chang, J. P.; Cheng, J.; and Danescu-Niculescu-Mizil, C. 2020. Don't Let Me Be Misunderstood:Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of The Web Conference 2020*, WWW '20, 2066–2077. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4743–4754. Hong Kong, China: Association for Computational Linguistics.

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

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829. Florence, Italy: Association for Computational Linguistics.

Crossley, S. A.; Kyle, K.; and McNamara, D. S. 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3): 803–821.

Dahiya, S.; Sharma, S.; Sahnan, D.; Goel, V.; Chouzenoux, É.; Elvira, V.; Majumdar, A.; Bandhakavi, A.; and Chakraborty, T. 2021. Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter. In Zhu, F.; Ooi, B. C.; and Miao, C., eds., *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, 2732–2742. ACM.

Davidson, S.; Sun, Q.; and Wojcieszak, M. 2020. Developing a New Classifier for Automated Identification of Incivility in Social Media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 95–101. Online: Association for Computational Linguistics.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, 512–515. AAAI Press.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Fancellu, F.; Lopez, A.; and Webber, B. 2016. Neural Networks For Negation Scope Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 495–504. Berlin, Germany: Association for Computational Linguistics.

Fanton, M.; Bonaldi, H.; Tekiroğlu, S. S.; and Guerini, M. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3226–3240. Online: Association for Computational Linguistics.

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

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4).

Garland, J.; Ghazi-Zahedi, K.; Young, J.-G.; Hébert-Dufresne, L.; and Galesic, M. 2022. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1): 3.

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

Ghosh, D.; Vajpayee, A.; and Muresan, S. 2020. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, 1–11. Online: Association for Computational Linguistics.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1336–1350. Online: Association for Computational Linguistics.

Hangartner, D.; Gennaro, G.; Alasiri, S.; Bahrich, N.; Bornhoft, A.; Boucher, J.; Demirci, B. B.; Derksen, L.; Hall, A.; Jochum, M.; et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50): e2116310118.

Jo, Y.; Bang, S.; Manzoor, E.; Hovy, E.; and Reed, C. 2020. Detecting Attackable Sentences in Arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1–23. Online: Association for Computational Linguistics.

Lambert, C.; Rajagopal, A.; and Chandrasekharan, E. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 548–559.

Levy, S.; Kraut, R. E.; Yu, J. A.; Altenburger, K. M.; and Wang, Y.-C. 2022. Understanding conflicts in online conversations. In *Proceedings of the ACM Web Conference 2022*, 2592–2602.

Liu, P.; Guberman, J.; Hemphill, L.; and Culotta, A. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, 181–190. AAAI Press.

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

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. In Pfeffer, J.; Budak, C.; Lin, Y.; and Morstatter, F., eds., *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, 369–380. AAAI Press.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153–157.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Munger, K. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3): 629–649.

Niculae, V.; Kumar, S.; Boyd-Graber, J.; and Danescu-Niculescu-Mizil, C. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1650–1659. Beijing, China: Association for Computational Linguistics.

Nobata, C.; Tetreault, J. R.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive Language Detection in Online User Content. In Bourdeau, J.; Hendler, J.; Nkambou, R.; Horrocks, I.; and Zhao, B. Y., eds., *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 145–153. ACM.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Potash, P.; and Rumshisky, A. 2017. Towards Debate Automation: a Recurrent Model for Predicting Debate Winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2465–2475. Copenhagen, Denmark: Association for Computational Linguistics.

Pougué-Biyong, J.; Semenova, V.; Matton, A.; Han, R.; Kim, A.; Lambiotte, R.; and Farmer, D. 2021. DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Procter, R.; Webb, H.; Burnap, P.; Housley, W.; Edwards, A.; Williams, M. L.; and Jirotka, M. 2019. A Study of Cyber Hate on Twitter with Implications for Social Media Governance Strategies. In Liakata, M.; and Vlachos, A., eds., *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019*.

Pruksachatkun, Y.; Yeres, P.; Liu, H.; Phang, J.; Htut, P. M.; Wang, A.; Tenney, I.; and Bowman, S. R. 2020. jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 109–117. Online: Association for Computational Linguistics.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 2019. A Benchmark Dataset for Learning to Intervene in

Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4755–4764. Hong Kong, China: Association for Computational Linguistics.

Richards, R. D.; and Calvert, C. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, 553.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. Vancouver, Canada: Association for Computational Linguistics.

Sadeque, F.; Rains, S.; Shmargad, Y.; Kenski, K.; Coe, K.; and Bethard, S. 2019. Incivility Detection in Online Comments. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 283–291. Minneapolis, Minnesota: Association for Computational Linguistics.

Schieb, C.; and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Schmidt, A.; and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. Valencia, Spain: Association for Computational Linguistics.

Shnarch, E.; Alzate, C.; Dankin, L.; Gleize, M.; Hou, Y.; Choshen, L.; Aharonov, R.; and Slonim, N. 2018. Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 599–605. Melbourne, Australia: Association for Computational Linguistics.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 613–624. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450341431.

Tekiroğlu, S. S.; Chung, Y.-L.; and Guerini, M. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1177–1190. Online: Association for Computational Linguistics.

Tsagkias, M.; Weerkamp, W.; and De Rijke, M. 2009. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 1765–1768.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, 80–93. Florence, Italy: Association for Computational Linguistics.

Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2289–2303. Online: Association for Computational Linguistics.

Wachs, S.; Krause, N.; Wright, M. F.; and Gámez-Guadix, M. 2023. Effects of the Prevention Program "HateLess. Together against Hatred" on Adolescents' Empathy, Self-efficacy, and Countering Hate Speech. *Journal of youth and adolescence*, 52(6): 1115–1128.

Weld, G.; Zhang, A. X.; and Althoff, T. 2021. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit.

Weld, G.; Zhang, A. X.; and Althoff, T. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1): 1121–1132.

Wiegand, M.; Ruppenhofer, J.; and Kleinbauer, T. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 602–608. Minneapolis, Minnesota: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Yano, T.; and Smith, N. 2010. What's Worthy of Comment? Content and Comment Volume in Political Blogs. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1): 359–362.

Yu, X.; Blanco, E.; and Hong, L. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5918–5930. Seattle, United States: Association for Computational Linguistics.

Yuan, J.; and Singh, M. P. 2023. Conversation Modeling to Predict Derailment. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1): 926–935.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.

Checklist

1. For most authors...

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

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

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Measuring Conversation Incivility, A Corpus of [...] Incivility, Corpus Analysis, and Experiments

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Measuring Conversation Incivility, A Corpus of [...] Incivility, Corpus Analysis, and Experiments

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we include the results in the URL in the Introduction

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

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

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

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

(a) If your work uses existing assets, did you cite the creators? Yes, see the A Corpus of [...] Incivility, Corpus Analysis, and Experiments

(b) Did you mention the license of the assets? Yes, see the Ethical Considerations

(c) Did you include any new assets in the supplemental material or as a URL? No

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

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

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see the A Corpus of [...] Incivility

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, we have and will release it

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

(a) Did you include the full text of instructions given to participants and screenshots? Yes, see the A Corpus of [...] Incivility

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see the Ethical Considerations

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see the Ethical Considerations

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see the Ethical Considerations