Measuring and Forecasting Conversation Incivility: the Role of Antisocial and Prosocial Behaviors

Xinchen Yu[1], **Hayden Arnold**[2], **Benjamin Su**[3], **Eduardo Blanco**[1]
[1]University of Arizona
[2]Korea Advanced Institute of Science and Technology
[3]University of Southern California
{xinchenyu, eduardoblanco}@arizona.edu, harnold@kaist.ac.kr, subenjam@usc.edu

Abstract

This paper focuses on the task of measuring and forecasting incivility in conversations following replies to hate speech. Identifying replies that steer conversations away from hatred and elicit civil follow-up conversations sheds light into effective strategies to engage with hate speech and proactively avoid further escalation. We propose new metrics that take into account various dimensions of antisocial and prosocial behaviors to measure the conversation incivility following replies to hate speech. Our best metric aligns with human perceptions better than prior work. Additionally, we present analyses on a) the language of antisocial and prosocial posts, b) the relationship between antisocial or prosocial posts and user interactions, and c) the language of replies to hate speech that elicit follow-up conversations with different incivility levels. We show that forecasting the incivility level of conversations following a reply to hate speech is a challenging task. We also present qualitative analyses to identify the most common errors made by our best model.

Introduction

Online discussion platforms enable new forms of interaction and have democratized public discourse on an immense scale. Hate speech, however, challenges their benefits. It harms individuals who suffer from personal attacks (Olteanu et al. 2018), distracts people from the goals of discussions (Arazy, Yeo, and Nov 2013), and even influences offline hate crimes (Farrell et al. 2019). Engaging with hate speech, for example by rebutting hateful content or discouraging hateful speakers, has emerged as a promising approach to address this problem. Compared with reactive moderation in which "bad actors" and "objective content" are identified and removed (Chang, Schluger, and Danescu-Niculescu-Mizil 2022), properly engaging with hate speech may divert the discourse away from hatred and avoid escalating tense situations. Further, engagement does not restrict free speech (Schieb and Preuss 2016).

Prior work has focused on modeling replies to hate speech, including corpora construction (Mathew et al. 2019; Chung et al. 2019), fine-grained categorization (Mathew et al. 2019; Yu et al. 2023), and generation (Zhu and Bhat 2021; Gupta et al. 2023; Chung and Bright 2024). Still, the effectiveness of replies to hate speech is understudied. When engaging with hate speech, conversational outcomes vary depending on the strategies used. Consider the Reddit conversation in Figure 1. There are three direct replies to the hateful post: two from actual Reddit users and one from a generative model, ChatGPT (OpenAI 2024). The first reply disagrees with the hateful post showing empathy and support—two prosocial behaviors. It successfully steers the author of the hateful post towards civil behavior in the follow-up conversation. The second reply uses denigrating language—an antisocial behavior—and results in an additional uncivil post. The third reply is polite but not as rich as genuine replies written by Reddit users. The effectiveness of such generic replies remains unclear.

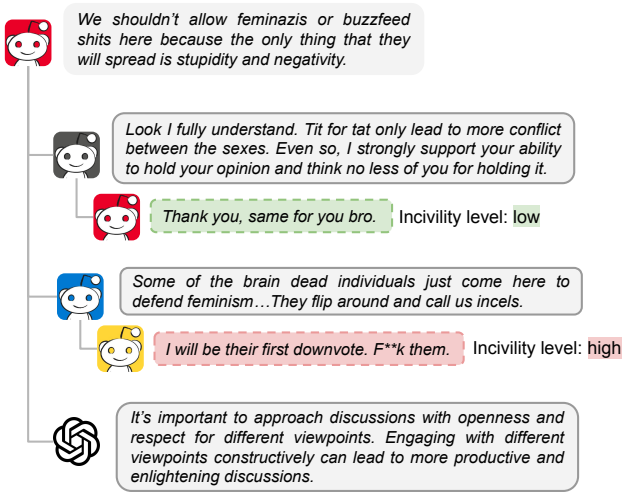

Figure 1: Hateful Reddit post (top), three direct replies, and the follow-up conversations. The first reply steers the follow-up conversation towards civil behaviors. The second reply elicits additional incivility. The last reply is generated by ChatGPT. It does not address the hateful post directly or elicit a follow-up conversation.

In this paper, we forecast the incivility of conversations following replies to hate speech in user-generated content. Doing so opens the door to identify content that is likely—and unlikely—to result in escalation of hatred before this

undesirable outcome becomes a reality. The forecasting proposed here could be used to direct moderation efforts to conversations where human intervention is most needed to reduce hatred. At the same time, the forecasting would also minimize moderation when users' replies to hateful posts organically elicit civil conversations. Beyond real-world applications, our work provides data-driven observations into the language most effective at discouraging incivility as well as strategies to navigate difficult conversations.

We take a broad view of conversation incivility by considering a combination of antisocial, prosocial, and neutral behaviors. Antisocial behaviors have been defined as acts intended to harm or disadvantage another individual (Sage, Kavussanu, and Duda 2006), whereas prosocial behaviors are acts intended to help or benefit another person (Eisenberg, Fabes, and Damon 1998). Different from traditional approaches that model antisocial or prosocial behaviors separately (Zhang et al. 2018; Bao et al. 2021; Lambert, Rajagopal, and Chandrasekharan 2022), we make the first attempt to jointly model both. Inspired by the fact that prosocial behaviors are the opposite of antisocial behaviors (Bar-Tal 1976) but not a dichotomy, we examine whether their presence in a conversation is effective at estimating conversation incivility. Notably, instead of modeling a single dimension, for example only considering norm violations for antisocial behaviors, we work with several dimensions. As concepts of antisocial and prosocial behaviors are complex (Fortuna and Nunes 2018; Bao et al. 2021), we argue that including different dimensions could capture the nuances and help us understand which aspects contribute most to the perceptions of conversation incivility.

This study draws its conclusions from a large Reddit dataset. We start by studying the conversations following replies to hate speech from three complementary perspectives: (a) presence of prosocial and antisocial behaviors discussed in the literature, (b) linguistic features of prosocial and antisocial behaviors, and (c) user interactions. Then, we analyze the role of antisocial and prosocial behaviors in measuring conversation incivility. Human validation demonstrates that combining several dimensions of both kinds of behaviors results in a more robust metric than prior work. Finally, we experiment with classifiers to forecast the incivility of the conversation following a reply to hateful content. The experimental results show that the task is challenging, and we close with an error analysis.

In summary, our main contributions are:

- New metrics that take into account several dimensions of both antisocial and prosocial behaviors in measuring conversation incivility following replies to hate speech;
- Comparing different types of replies to hate speech with respect to language usage, including a) the difference between antisocial replies and prosocial replies and b) the difference between replies eliciting conversations with different incivility levels: high, medium and low;
- Analyzing antisocial and prosocial behaviors in user interactions with regard to two scenarios: re-engagement and multi-turn conversations;
- Building models to predict conversation incivility level and presenting a qualitative error analysis.

Related Work

Antisocial Behavior Prior work has studied a wide range of antisocial behaviors in online platforms such as Reddit (Vidgen et al. 2021), Instagram (Liu et al. 2018), and Wikipedia (Zhang et al. 2018). These antisocial behaviors include hate speech (Röttger et al. 2021), abusive language (Vidgen et al. 2021), offensive language (Zampieri et al. 2019), toxicity (Pavlopoulos et al. 2020), and norm violations (Lambert, Rajagopal, and Chandrasekharan 2022). Most of these previous studies focus on reactive moderation, that is, detecting antisocial behaviors after they have occurred. Our work, similar to another line of prior research (Zhang et al. 2018; Lambert, Rajagopal, and Chandrasekharan 2022), aims at forecasting whether (future) conversations will be uncivil. We focus on forecasting incivility of conversations following replies to hateful posts.

Prosocial Behavior Existing work on prosocial behaviors has explored politeness (Danescu-Niculescu-Mizil et al. 2013), empathy (Buechel et al. 2018; Liu et al. 2024), sympathy and encouragement (Sosea and Caragea 2022), positiveness (Ziems et al. 2022), donation (Dong, Xu, and Mihalcea 2022) and norms (Ziems et al. 2023). Specifically, Bao et al. (2021) and Lambert, Rajagopal, and Chandrasekharan (2022) have proposed various prosocial metrics to quantify and predict prosocial outcomes in follow-up conversations. Instead of using hand-crafted lexicon features, we leverage deep neural networks to identify prosocial behaviors. To the best of our knowledge, we are the first to examine whether prosocial and antisocial behaviors even out the incivility people perceive in a conversation.

Forecasting Conversational Incivility Several efforts have predicted whether an event will lead to derailing future conversations. This includes predicting future removal of a post (Cheng et al. 2017), personal attacks (Zhang et al. 2018), and incivility intensity (Liu et al. 2018; Dahiya et al. 2021; Yu, Blanco, and Hong 2024). Given a conversation, prior work uses different metrics to estimate incivility. Liu et al. (2018) considers the number of uncivil posts. Dahiya et al. (2021) averages the scores of comments from a toxicity classifier. Yu, Blanco, and Hong (2024) take into account the number of uncivil and civil comments as well as the unique users involved in the conversation. This paper extends these works by combining several dimensions of both antisocial and prosocial comments for the first time. Doing so results in a metric that more closely resembles human judgments.

An Analysis of Reddit Conversations Following Replies to Hateful Posts

We first describe (a) the procedure to collect a large collection of relevant Reddit conversations (hateful post, reply, and follow-up conversation) and (b) several dimensions of antisocial and prosocial behaviors from the literature. Then, to better understand conversations following replies to hateful posts, we conduct analyses on such conversations containing prosocial and antisocial behaviors. Specifically, we investigate (a) linguistic characteristics to reveal differences in language use and (b) user interactions (e.g., re-engagement, multi-turn conversations).

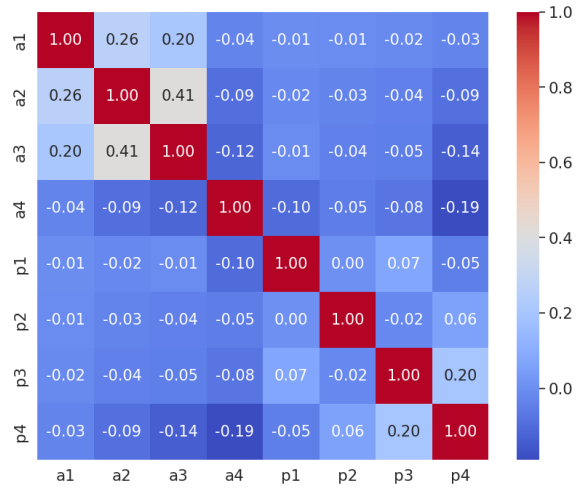

Figure 2: Spearman's rank correlation coefficients between antisocial and prosocial behaviors. a_i and p_i indicate antisocial and prosocial behaviors in the order described in the paper. All coefficients are low (< 0.3) except the one between a_2 (explicit hate speech) and a_3 (abusive language).

	p-value
Textual factors	
First pronoun	↑↑↑
Second pronoun	↓↓↓
Total tokens	↑↑↑
Negation cues	↑↑↑
Question mark	↑↑↑
Sentiment factors	
Disgust words	↑↑↑
Sadness words	↑↑↑
Negative words	↑↑↑
Positive words	↓↓↓
Happiness words	↓↓↓
Gratitude words	↓↓↓
Hostile words	↑↑↑
Angry words	↑↑↑

Table 1: Linguistic analysis comparing antisocial and prosocial posts. Three arrows indicates p-value < 0.001 (unpaired t-test). Arrow direction indicates whether higher values correlate with antisocial posts (up) or prosocial posts (down).

Dataset Our starting point is the Reddit dataset collected by Yu, Blanco, and Hong (2024). It consists of 1,382,596 posts from 39 subreddits and binary labels indicating whether each post is hateful. Reddit, known for its large size of user populations and diverse topics (Baumgartner et al. 2020), is an ideal source for studying online conversations. We first select all the hateful posts and create (*hateful post*, *reply*, *follow-up conversation*) triples by pairing (a) each hateful post with each of its direct replies and (b) each direct reply to the hateful post with the follow-up conversation. This strategy results in 25,225 hateful posts, 41,727 replies to the hateful posts, and a total of 124,172 posts in the follow-up conversations. After removing replies that have been moderated, we obtain 38,041 triples. Each triple consists of one hateful post, one reply, and the follow-up conversation. Follow-up conversations can be empty and the maximum length is 975 posts.

Identifying Antisocial and Prosocial Comments We work with four antisocial behaviors and four prosocial behaviors drawn from the literature. Our selection criteria includes not only theoretical background, but also whether corpora exist with which to train classifiers for large-scale automatic annotation. For antisocial behaviors, we work with the following: (a) offensive language: offensive or not offensive (Davidson et al. 2017), (b) explicit hate speech: hateful or not hateful (Qian et al. 2019), (c) abusive language: abusive or not abusive (Vidgen et al. 2021), and (d) norm violations (whether the content of a post is "*deleted*" or "*removed*"): yes or no. For prosocial behaviors, we work with the following: (a) empathy: expressing (including both weak and strong expressions) or not expressing (Sharma et al. 2020), (b) norms: expected or not expected (Ziems et al. 2023), (c) positiveness: positive or not positive (Ziems et al. 2022), and (d) politeness: polite or not polite (Danescu-Niculescu-Mizil et al. 2013).

We fine-tune a RoBERTa transformer (Liu et al. 2019) for each dimension of antisocial and prosocial behavior in order to build a classifier. We refer the reader the Appendices for the experimental results. Then, we use the resulting classifiers to indicate whether each post in the conversation following a reply to hate speech exhibits each antisocial and prosocial behavior.

A perhaps obvious question is whether the four antisocial and prosocial behaviors account for the same phenomena and thus differ only in name. We refer the reader to the original papers for details, but the answer is that that is not the case. Figure 2 shows the correlation coefficients between all antisocial and prosocial behaviors considered as observed in the follow-up conversations to replies to hateful posts. All the correlations are low (< 0.3) except the one between *explicit hate speech* and *abusive language*. In other words, each antisocial and prosocial behavior captures a distinct characteristic in a Reddit post.

The Language of Antisocial and Prosocial Posts

While it is well-known that prosocial behaviors are the opposite of antisocial behaviors (Bar-Tal 1976), the actual differences between these behaviors remains unknown. To find out the language differences, we compare linguistic features of antisocial and prosocial posts in the conversations following replies to hate. We consider a post antisocial if it is identified as such by any antisocial classifier, and similarly, we consider a post prosocial if it is identified as such by any prosocial classifier. We exclude posts identified as norm violations in this analysis, as their content is always "*deleted*." This results in 25,312 antisocial post and 25,260 prosocial posts. We consider both textual features (e.g., first and second pronouns, negation cues) and sentiment features (Crossley, Kyle, and McNamara 2017). Negation cues checks for presence in the list by Fancellu, Lopez, and Webber (2016).

We also conduct the Bonferroni correction as multiple hypothesis tests are performed.

Table 1 shows the results that passed the Bonferroni correction. Regarding textual features, people use significantly ($p < 0.001$) more first pronouns, tokens, negation cues and question marks when attacking and harming other individuals (antisocial behavior). There are more second pronouns (e.g., *you* and *your*) in prosocial posts, indicating caring behaviors (Hoffman 1996). Regarding sentiment features, there are significantly ($p < 0.001$) more negative, disgust, hostile and angry words in antisocial comments. On the other hand, we find more happiness, gratitude, and positiveness in prosocial comments.

Do Antisocial Behaviors and Prosocial Behaviors Influence User Interactions?

To understand whether antisocial and prosocial behaviors affect how people interact (e.g., re-engagement, multi-turn conversations), we analyze the structure of the follow-up conversations after replies to hateful posts. We work with all the posts in the follow-up conversations and answer the following questions.

Does Re-Engagement Differ after Receiving Antisocial or Prosocial Posts? Yes, it does. We focus on users who receive both antisocial and prosocial posts (4,537) and compare their re-engagement in the subsequent conversations. We analyze re-engagement from two perspectives: how often they re-enter after each antisocial (or prosocial) post (a) anywhere in the subsequent conversation and (b) immediately after the antisocial (or prosocial) post (i.e., direct replies). Results show that when receiving antisocial posts, the percentage of re-engagement—including both anywhere and immediately after the antisocial post—is significantly higher than when receiving prosocial comments (paired sample t-test, $p < 0.001$).

Are There More Antisocial Posts than Prosocial Posts in Multi-Turn Conversations? No, there are not. We consider multi-turn conversations those in which there are at least two turns between each pair of users in a conversation. We conduct paired sample t-tests to see which behaviors (antisocial or prosocial) are more frequent in these conversations. Results show that prosocial posts are significantly more frequent than antisocial posts ($p < 0.001$), despite the percentage of antisocial posts (20.38%) is slightly higher than prosocial posts (20.34%) in all the follow-up conversations. In other words, the amount of antisocial posts is roughly the same, but there are significantly more prosocial posts in multi-turn conversations—the longer the conversation, the better the tone.

Do People Who Engage in a Multi-Turn Conversation Display the Same Behavior? No, they do not. We examine whether two people in a multi-turn conversation behave similarly. That is, whether they both display either antisocial or prosocial behavior or a mix. For each pair of users in a multi-turn conversation, we calculate the percentage of antisocial and prosocial posts out of all posts directed at each other. Then, we calculate the difference in percentages of antisocial (*anti_diff*) and prosocial (*pro_diff*) per pair of users. Finally, we calculate the difference between *anti_diff* and *pro_diff* per pair. If both users have similar behaviors, the final difference will be small. On the other hand, if a user displays more antisocial (or prosocial) behaviors than the other, the absolute final difference would be large. Results show that there are significant differences (one-sample t-test, $p < 0.001$) in the behaviors of individuals participating in multi-turn conversations. In other words, it is common to have conversations in which two users display opposite behaviors—antisocial and prosocial behaviors do not elicit the same behavior from the second speaker.

Measuring Conversation Incivility

We aim to define an automatic metric to assess conversation incivility so that low-cost, large-scale annotations become a reality. Unlike previous work which considers either *one* antisocial or *one* prosocial behavior, our proposal considers (a) both antisocial and prosocial behaviors as well as neutral behaviors and (b) four types of antisocial and prosocial behaviors. By experimenting with several weighting mechanisms, we replicate all previous metrics. Careful evaluation allows us to conclude that our metric outperforms previous proposals. In fact, when our automated metric is considered an "annotator" to identify which of two follow-up conversations is more uncivil, it obtains inter-annotator agreement (compared to a "real" (human) annotator) above the threshold to be considered reliable.

A Metric to Measure Conversation Incivility Our metric includes three main components accounting for antisocial, prosocial, and neutral behavior. Further, we include user re-entry behaviors (Backstrom et al. 2013) to identify whether the same author engages in multiple posts with the same behavior. Intuitively, we give less weight to the same author displaying the same behavior in multiple posts compared to the same amount of posts by several authors.

Given a reply r to a hateful post, the conversation incivility score of the follow-up conversation ($S(r)$) consists of three components: antisocial component $A(r)$, prosocial component $P(r)$, and neutral component $N(r)$. The antisocial component consists of up to the four dimensions introduced earlier: a_1 (offensive language), a_2 (explicit hate speech), a_3 (abusive language), and a_4 (norm violations). Similarly, the prosocial component also consists of up to four dimensions: p_1 (empathy), p_2 (norms), p_3 (positiveness), and p_4 (politeness). The neutral component includes posts that display neither antisocial nor prosocial behaviors. We assign equal weights to the antisocial and prosocial dimensions. For example, if we only consider *offensive language* (a_1) and *abusive language* (a_3) as indicators of incivility, a weight 0.5 is assigned to each of the two dimensions.

In addition to the amount of posts displaying the antisocial and prosocial behaviors, user re-entry is an important factor in measuring incivility (Backstrom et al. 2013). As pointed out by Yu, Blanco, and Hong (2024) within the hate and counterhate domain, the same amount of posts displaying a behavior should be given more weight if they come from several users as opposed to a single, prolific user. We adapt

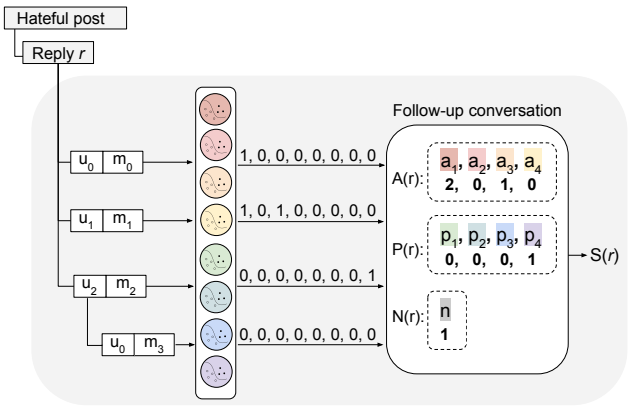

Figure 3: Illustration of our metric to estimate incivility of the conversation following a reply r to a hateful post ($S(r)$). We calculate four antisocial and four prosocial behaviors for each post. The metric consists of components to account for antisocial ($A(r)$), prosocial ($P(r)$), and neutral behaviors ($N(r)$); and considers not only the amount of each behavior but also whether different authors generate posts with the same behavior (not shown).

their insight to account for several prosocial and antisocial dimensions. Specifically, for each user u ($u = 0, 1, ..., k$), we calculate the number of antisocial and prosocial posts $c_{a_i,u}$ and $c_{p_i,u}$ per dimension i ($i = 1, 2, 3, 4$) as well as the count of neutral comments $c_{n,u}$. A function f (e.g., square root) is applied on $c_{a_i,u}$, $c_{p_i,u}$, and $c_{n,u}$ to differentiate instances in which the same user makes several posts with the same behavior (e.g., with $f = \sqrt{}$, 10 antisocial posts by 10 users are considered as uncivil as 100 posts by the same person). The choice of f is not crucial, as the metric is better suited to be used in relative terms (i.e., for comparison purposes) rather than absolute terms. Summing up over all users in the conversation, we have: $A(r) = \sum_{u=0}^{k} f(c_{a_i,u})$, $P(r) = \sum_{u=0}^{k} f(c_{p_i,u})$, and $N(r) = \sum_{u=0}^{k} f(c_{n,u})$.

The more antisocial posts, the more uncivil a conversation is. Similarly, the more prosocial and neutral posts, the more civil a conversation is. We thus define $S(r)$ as follows:

$$S(r) = \alpha \cdot A(r) - \beta \cdot P(r) - (1 - \alpha - \beta) \cdot N(r)$$

Parameters α and β are weights for the antisocial and prosocial components (and the neutral component). Figure 3 illustrates the procedure to calculate $S(r)$.

The specific antisocial and prosocial dimensions as well as α and β are parameters that must be tuned. We choose $\alpha, \beta \in [0, 1]$ with $\alpha + \beta \leq 1$. We experiment with 0.05 increments and compare against human judgments as explained below. Our data-driven metric is the first to consider both antisocial and prosocial behaviors, and, importantly, several dimensions of each. Note that our approach subsumes all prior metrics to measure conversation incivility. By trying all combinations of antisocial and prosocial behaviors, we subsume previous efforts considering either one antisocial or prosocial behavior (Liu et al. 2018). We also subsume efforts considering either antisocial ($\alpha = 1$ and $\beta = 0$) or prosocial ($\alpha = 0$ and $\beta = 1$) behaviors (Bao et al. 2021), antisocial and non-antisocial behaviors ($\alpha \in (0, 1)$ and $\beta = (1 - \alpha)/2$), and prosocial and non-prosocial behaviors ($\beta \in (0, 1)$ and $\alpha = (1 - \beta)/2$) (Lambert, Rajagopal, and Chandrasekharan 2022).

Tuning the Metric Our metric definition must be tuned to determine the optimal antisocial and prosocial behaviors as well as α and β. While one could make decisions inspired by the literature, we argue that a data-driven approach is more sound. To this end, we conduct human annotations so that we can identify the optimal choices by comparing how closely the metric estimates human assessments.

First, we randomly select two (hateful post, reply, follow-up conversations) triples accounting for a variety of lengths in the follow-up conversations. Specifically, we divide follow-up conversations into short (≤ 5 posts), medium (> 5 posts and ≤ 10 posts) and long (> 10 posts). Then, we generate 40 pairs of triples where the follow-up conversation belong to the following lengths: *short* and *short*, *short* and *medium*, *short* and *long*, *medium* and *medium*, *medium* and *long*, and *long* and *long*. Finally, we shuffle the triples in the pair. The first step ensures that we have a representative sample of follow-up conversations as far as length. Thus, the tuning process is designed to result in a robust metric regardless of the length of the follow-up conversation.

Second, we employ two native English speakers to manually annotate which follow-up conversation in the triple is more uncivil in a pair. The interface displays the full triple (hateful post, reply, and follow-up conversation), similar to the illustration shown in Figure 1. The Cohen's κ is 0.84 (90% accuracy) among the 240 pairs, which is considered nearly perfect (Artstein and Poesio 2008). We show the Cohen's κ of each group as follows: a) long vs. long: 0.90, b) long vs. medium: 0.89, c) long vs. short: 0.75, d) short vs. short: 0.74, e) short vs. medium: 0.79, and f) medium vs. medium: 0.85. Cohen's κ is lower when both follow-up conversations in a pair are short (i.e., short vs. short) or when their length varies substantially (i.e., long vs. short).

Third, we try all combinations of antisocial and prosocial behaviors as well as α and β to find the optimal choices. Armed with the manually annotated benchmark, this task is trivial. Note that any instantiation of our metric ($S(r)$) can be used to tell which of two conversations is more uncivil by comparing their incivility scores. Figure 4 provides the Cohen's κ coefficients comparing the results obtained with our metric and the human annotators. For each combination of antisocial and prosocial dimensions, the figure presents the highest κ after trying all combinations of α and β.

We make the following observations from the results:

- The best metric takes into account both antisocial (a_1, a_2, a_3) and prosocial behaviors (p_3) with $\alpha = 0.75$ and $\beta = 0.15$. It obtains a Cohen's κ of 0.68, which is considered *substantial* agreement (Artstein and Poesio 2008). While lower than the "true" inter-annotator agreement ($\kappa = 0.84$), our metric is considered a reliable annotator ($\kappa = 0.68$; coefficients between 0.6 and 0.8 are considered *substantial* agreement).
- Antisocial behaviors are much more informative than

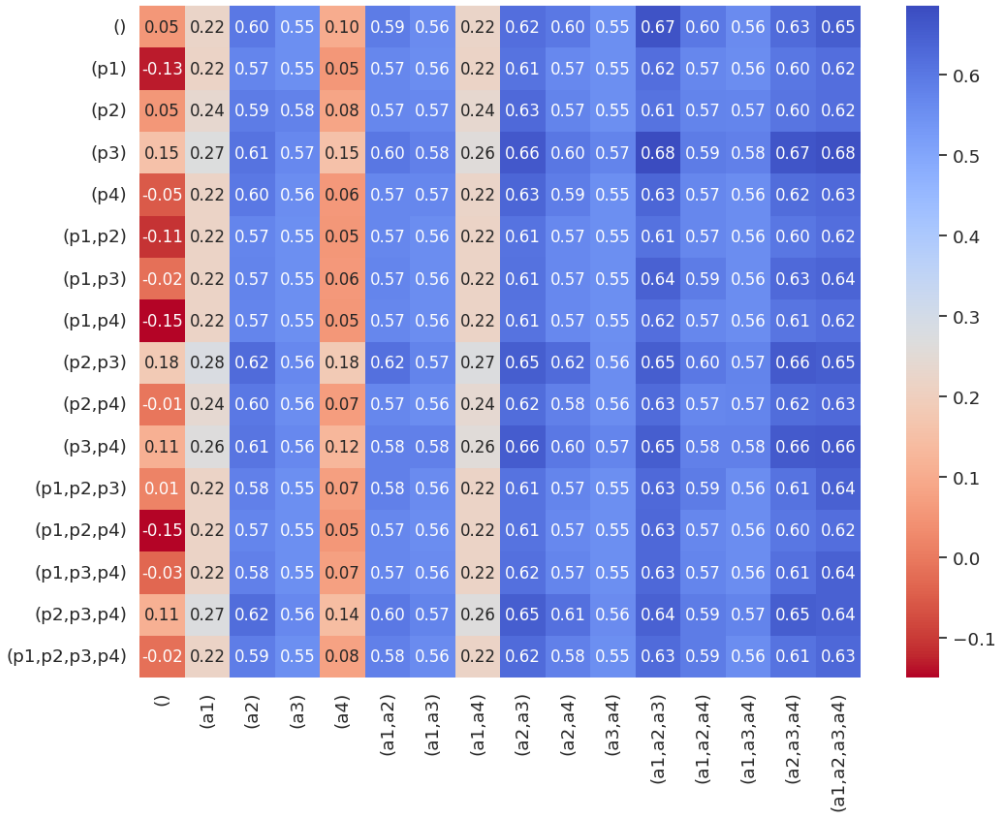

Figure 4: Cohen's κ coefficients between human annotations (after adjudication) and using our metric to determine which of two conversations is more uncivil. Cells indicate the highest κ obtained with the corresponding combination of antisocial and prosocial behaviors after trying all combinations of α and β. Prosocial behaviors by themselves underperform, and combining several antisocial behaviors outperforms individual antisocial behaviors. The optimal choice is (a_1, a_2, a_3) and (p_3).

prosocial behaviors. Indeed, considering only antisocial behaviors (top row) yields $\kappa = 0.67$.

- Considering several antisocial behaviors is beneficial. Indeed, single antisocial behavior yields $\kappa = 0.22 - 0.60$, while considering more than one, yields $\kappa = 0.67$.
- While prosocial behaviors obtain poor coefficients by themselves (first column), they are modestly beneficial when combined with antisocial behaviors.

Categorizing Conversation Incivility Having identified the best metric, using it to annotate the incivility score of all the follow-up conversations in our corpus is straightforward. We use the incivility scores to group all the replies to hateful posts based on the incivility quartiles in the follow-up conversations to the replies (top 25%, middle 50%, and bottom 25%). The exact score ranges are as follows:

- *Low* incivility: $S(r) \in (-16.38, -0.10]$;
- *Medium* incivility: $S(r) \in (-0.10, 0]$; and
- *High* incivility: $S(r) \in (0, 7.81]$.

Corpus Analysis

The final corpus consists of 38,041 (*hateful post*, reply, follow-up conversation) triples with the incivility scores of the follow-up conversations. 23.48% of the replies are followed by conversations with high incivility, 25.43% are followed by conversations with low incivility, and the rest (51.09%) are followed by conversations with medium incivility. We note that 92.10% of the replies with medium conversation incivility scores have no follow-up conversations. This motivates us to explore the replies to hateful posts from two perspectives. First, we investigate the language in the replies that is likely to elicit follow-up conversations regardless of incivility. Second, we explore the differences in language between follow-up conversations with high and low incivility. We therefore compare the linguistic differences between replies to hate speech that a) have and do not have follow-up conversations (high vs. medium, low vs. medium), and b) have different incivility levels of follow-up conversations when they are not empty (high vs. low). We run unpaired t-tests and report results in Table 2.

We observe several interesting findings:

- Regarding textual features, replies that use more personal pronouns (both 1st and 2nd person), tokens, negation cues and question marks are likely to attract future conversations to follow, and at the same time, more incivility in these follow-up conversations.
- Regarding sentiment factors, there are significantly more

		High vs Low		High vs Medium	Low vs Medium
		hateful post	reply	reply	reply
Textual factors	1st person pronouns	↑↑	↑↑↑	↑↑↑	↑↑↑
	2nd person pronouns	↑↑↑	↑↑↑	↑↑↑	↑↑↑
	Tokens	↑	↑↑↑	↑↑↑	↑↑↑
	Negation cues	↑↑↑	↑↑↑	↑↑↑	↑↑↑
	Quotations	↑↑	↑↑↑	↑↑↑	
	Question marks	↑↑↑	↑↑↑	↑↑↑	↑↑↑
Sentiment factors	Disgust words	↑↑↑	↑↑↑	↑↑↑	
	Sadness words	↑↑↑	↑↑↑	↑↑↑	
	Positive words	↓↓↓	↓↓↓	↓↓↓	↑↑↑
	Negative words	↑↑↑	↑↑↑	↑↑↑	
	Anger words		↑↑↑	↑↑↑	
	Hostile words	↑↑	↑↑↑	↑↑↑	

Table 2: Linguistic analysis comparing (a) hateful posts that result in high and low incivility in the follow-up conversations to their replies (Column 3) and (b) replies to hateful posts that result in high, medium and low incivility in the follow-up conversation (Columns 4–6). 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 the first group (up or the second group (down) in each pairwise comparison. Tests that pass the Bonferroni correction have background color.

negativeness, disgust, sadness, anger, and hostility in the replies followed by conversations with high incivility as well as less positiveness. Besides the replies, the hateful posts that contain more negativeness and less positiveness elicit more incivility in the follow-up conversations.

Experiments

We experiment with models to determine the incivility level (i.e., high, medium, or low) of the conversation following a reply to the hateful post. We randomly split the 38,041 instances as follows: 70% for training, 15% for validation and 15% for testing. To investigate whether taking into account the hateful posts is beneficial, we consider two textual inputs: a) the reply to the hateful post (reply), and b) the hateful post and the reply (hate + reply). We experiment with supervised approaches and prompting LLMs.

Supervised Approaches

We start with the off-the-shelf RoBERTa transformer (Liu et al. 2019) released by Hugging Face (Wolf et al. 2020). We also experiment with FLAN-T5 (Chung et al. 2024). As the performance is very close to RoBERTa, we detail the results with FLAN-T5 in the Appendices. We explore two strategies to improve performance of these base models.

Pretraining with Related Tasks Pretraining could be seen as a two-stage fine-tuning process. First, we fine-tune a RoBERTa-base classifier with a related corpus, and then with our own corpus. We use the following related corpora: hate speech (Davidson et al. 2017), counterspeech (Yu, Blanco, and Hong 2022), stance (Pougué-Biyong et al. 2021), and sentiment (Rosenthal, Farra, and Nakov 2017).

Blending Additional Data Blending (Shnarch et al. 2018) starts combining both a related corpus and our corpus in the fine-tuning process. It decreases the portion of instances from the related corpus after each epoch by a fixed ratio α. The last m epochs are trained with data only from our own corpus. The blending hyperparameters (α and m) are tuned like any other hyperparameter. We use the same corpora as above for pretraining purposes.

GPT-4o: Zero- and Few-Shot

Having witnessed the success of large language models and prompt engineering (Mishra et al. 2022), we are curious to see whether they outperform supervised approaches using substantially smaller models in our tasks. We experiment with GPT-4o using Microsoft Azure API. For few-shot prompts, surprisingly, GPT-4o obtains worse results with 4-shot than 1-shot prompting. We refer the reader the Appendices for additional details about the prompts and examples.

Experimental Results

Table 3 shows the results per label and weighted averages. We provide here results pretraining and blending with the most beneficial tasks: stance for pretraining and counterspeech for blending (optimal $\alpha = 1$). Regarding supervised approaches, using only the reply as input offers competitive performance with F1 scores up to 0.47 compared with the random baseline (F1: 0.47 vs. 0.35). Using both the hate comment and the reply as input yields better results (F1: 0.49 vs 0.47). Finally, the network that takes both the hate comment and the reply as input and blends the counterspeech corpus or is pretrained with the stance corpus yields the best results (F1: 0.51). GPT-4o, however, performs much worse than supervised approaches on our task (F1: 0.42 vs 0.51).

Error Analysis

While the supervised approaches outperform prompting GPT-4o, forecasting conversation incivility level is a challenging task. We manually analyze 100 randomly sampled errors made by our best model (*hate+reply+blending*) to reveal the most common error types (Table 4).

	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	0.52	1.00	0.68	0.00	0.00	0.00	0.27	0.52	0.35
RoBERTa, training with												
reply	0.38	0.52	0.44	0.60	0.70	0.64	0.33	0.10	0.15	0.48	0.51	0.47
+ pretraining†	0.45	0.38	0.41	0.58	0.83	**0.68**	0.34	0.09	0.14	0.49	0.54	0.49
+ blending†	0.43	0.49	0.46	0.60	0.73	0.66	0.35	0.15	0.21	0.50	0.53	0.50
hate + reply	0 .39	0.45	0.42	0.60	0.62	0.61	0.32	0.26	**0.29**	0.48	0.49	0.49
+ pretraining†‡	0.42	0.53	**0.47**	0.61	0.72	0.66	0.38	0.15	0.22	0.51	0.53	**0.51**
+ blending†‡	0.44	0.51	**0.47**	0.62	0.72	0.66	0.34	0.17	0.23	0.51	0.53	**0.51**
GPT-4o, prompting with												
reply, Zero-Shot	0.30	0.59	0.40	0.52	0.36	0.43	0.24	0.18	0.21	0.40	0.37	0.37
reply, Few-Shot	0.36	0.40	0.38	0.52	0.51	0.51	0.26	0.25	0.25	0.42	0.42	**0.42**
hate + reply, Zero-Shot	0.29	0.56	0.38	0.53	0.45	0.48	0.25	0.11	0.16	0.40	0.39	0.38
hate + reply, Few-Shot	0.33	0.42	0.37	0.50	0.47	0.48	0.26	0.24	0.25	0.40	0.40	0.40

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 + reply* model. Training with the *hate + reply* coupled with pretraining with stance or both pretraining or blending counterspeech yields the best results (F1: 0.51).

Error Type	%	Example	Prediction	Gold
Rhetorical question	26	Hate: *I just enjoy calling out retards on the internet, it's not a crime.* Reply: *Maybe go look in a mirror?*	low	high
Irony or sarcasm	18	Hate: *That's stupid and you're stupid.* Reply: *I'd make a rebuttal but I hate getting into fights with children.*	low	high
Swear words in the reply	13	Hate: *Saying all this during Ramadan too? You're a shitty Muslim.* Reply: *From people like you who find it normal to abuse us. You are a hypocrite, pathetic, just another sheep. You are a shitty human.*	high	low
Request	8	Hate: *Lmao when you have straight make shit up is when people know you're full of shit.* Reply: *What was made up? Can you articulate your point?*	low	medium
Negation	5	Hate: *He was literally dying, and you motherf**kers laughed at it.* Reply: *You seem to imply he was a good person. Doesn't make the jokes any better, but the dude wasn't a role model.*	high	medium

Table 4: Most common error types made by the best model with predictions made by *hate + reply + blending*.

First, 26% of the errors contain rhetorical questions. In the example, the model fails to interpret that the reply is not expecting a reply. Instead, the rhetorical question is used to attack the author of the hateful post. Irony or sarcasm are present in 18% of the errors. The reply in the example uses irony (e.g., calling people "*children*") to humiliate the author of the hateful post, resulting in more incivility in the follow-up conversation (not shown).

More concerningly, some replies contain swear words but they do not elicit additional incivility (13% of errors). The reply states that the author of the hateful post is a "shitty human," and the model considers this would attract more conflicts. However, the follow-up conversation has low incivility. We also found the model struggles with replies questioning the validity of the content in the hateful post or asking for more evidence (Walton 2005). Finally, the model errs when there are negations in the replies. This accounts for 5% of the errors. The reply in the example contains multiple negation cues and the model mislabels it.

Conclusion and Discussion

In this paper, we work on the task of forecasting incivility of the conversations following replies to hate speech. We have presented new metrics that take into account several dimensions of antisocial and prosocial behaviors in measuring conversation incivility. The validation on a human-annotated benchmark demonstrates that it is worth accounting for both antisocial and prosocial behaviors, although the latter play a smaller role. Crucially, our metric aligns with human perceptions more closely than prior work modeling either prosocial or antisocial behaviors, or a single dimension of each kind of behavior. We find offensive language, explicit hate speech, and abusive language are useful in representing antisocial behaviors. Our extensive analy-

ses unpack one of the potential reasons: some conversations may include prosocial behaviors, yet people behave differently (e.g., one person misbehaves regardless of receiving prosocial comments from the others), therefore the overall incivility people perceive is still very high.

Experimental results show that supervised methods outperform prompting LLMs. Specifically, taking into account the hateful posts as well as blending or pretraining with additional corpora yield improvements, yet forecasting future incivility of the conversation following replies to hate speech is still a challenging task.

Limitations Since we do not conduct randomized controlled experiments, our linguistic analyses should not be interpreted as causal statements. Our metrics do not consider the dynamic and multi-layered structure of follow-up conversations. Finally, we identify antisocial and prosocial behaviors with classifiers. They obtain good results but are not perfect. Due to class imbalance in the training data, some classifiers tend to be biased towards the majority class (e.g., not abusive). Future research could explore the impact of classifier performance on the conversation incivility metric.

Ethical Statement

Reddit data are public available. We recognized the public nature of this information does not imply users' consent or willingness to share their data (Fiesler and Proferes 2018). We made the following efforts in protecting personal information of the Reddit community: First, we obfuscate user names to avoid identification of specific users. Second, we only report incivility scores of the follow-up conversations after replies to hate speech and do not publish these follow-up conversations. Third, upon release of our corpus, we will only release the IDs and labels of replies along with the hateful posts. Finally, in compliance with Reddit's policy, we would like to make sure that our dataset will be reused for non-commercial research only.[1]

The annotators were warned of the potential uncivil content before working on our task. They were also informed to end the annotation whenever feel uncomfortable or frustrated. We provide annotators with access to supporting services throughout the task. We acknowledge the risk associated with releasing the corpus, yet we believe the benefit of bringing to light what replies could mitigate future incivility outweighs any risks associated with the corpus release.

References

Arazy, O.; Yeo, L.; and Nov, O. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology*, 64(8): 1634–1648.

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

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*, 13–22.

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*, 1134–1145.

Bar-Tal, D. 1976. Prosocial behavior: Theory and research.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Buechel, S.; Buffone, A.; Slaff, B.; Ungar, L.; and Sedoc, J. 2018. Modeling Empathy and Distress in Reaction to News Stories. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4758–4765. Brussels, Belgium: Association for Computational Linguistics.

Chang, J. P.; Schluger, C.; and Danescu-Niculescu-Mizil, C. 2022. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–37.

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*, 1217–1230.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Chung, Y.-L.; and Bright, J. 2024. On the Effectiveness of Adversarial Robustness for Abuse Mitigation with Counterspeech. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6988–7002. Mexico City, Mexico: Association for Computational Linguistics.

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 Korhonen, A.; Traum, D.; and Màrquez, L., eds., *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: 803–821.

Dahiya, S.; Sharma, S.; Sahnan, D.; Goel, V.; Chouzenoux, E.; Elvira, V.; Majumdar, A.; Bandhakavi, A.; and Chakraborty, T. 2021. Would your tweet invoke hate on

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

the fly? forecasting hate intensity of reply threads on twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2732–2742.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. In Schuetze, H.; Fung, P.; and Poesio, M., eds., *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 250–259. Sofia, Bulgaria: Association for Computational Linguistics.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.

Dong, M.; Xu, X.; and Mihalcea, R. 2022. Text-Aware Graph Embeddings for Donation Behavior Prediction. In Ustalov, D.; Gao, Y.; Panchenko, A.; Valentino, M.; Thayaparan, M.; Nguyen, T. H.; Penn, G.; Ramesh, A.; and Jana, A., eds., *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, 60–69. Gyeongju, Republic of Korea: Association for Computational Linguistics.

Eisenberg, N.; Fabes, R. A.; and Damon, W. 1998. Handbook of child psychology. *Social, emotional*.

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

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, 87–96. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362023.

Fiesler, C.; and Proferes, N. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, 4(1): 2056305118763366.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2024-08-06.

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

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.

Gupta, R.; Desai, S.; Goel, M.; Bandhakavi, A.; Chakraborty, T.; and Akhtar, M. S. 2023. Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5792–5809. Toronto, Canada: Association for Computational Linguistics.

Hoffman, M. L. 1996. Empathy and moral development. *The annual report of educational psychology in Japan*, 35: 157–162.

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.

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 International AAAI Conference on Web and Social Media*, volume 12.

Liu, T.; Ungar, L.; Kording, K.; and McGuire, M. 2024. Measuring Causal Effects of Civil Communication without Randomization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 958–971.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: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 *Proceedings of the international AAAI conference on web and social media*, volume 13, 369–380.

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

Mishra, S.; Khashabi, D.; Baral, C.; Choi, Y.; and Hajishirzi, H. 2022. Reframing Instructional Prompts to GPTk's Language. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 589–612. Dublin, Ireland: Association for Computational Linguistics.

Olteanu, A.; Castillo, C.; Boy, J.; and Varshney, K. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

OpenAI. 2024. ChatGPT. https://www.openai.com/chatgpt. Accessed: 2024-08-06.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305. Online: 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)*.

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 Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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.

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

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 41–58. Online: Association for Computational Linguistics.

Sage, L.; Kavussanu, M.; and Duda, J. 2006. Goal orientations and moral identity as predictors of prosocial and antisocial functioning in male association football players. *Journal of Sports Sciences*, 24(05): 455–466.

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.

Sharma, A.; Miner, A.; Atkins, D.; and Althoff, T. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276. Online: 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 Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 599–605. Melbourne, Australia: Association for Computational Linguistics.

Sosea, T.; and Caragea, C. 2022. EnsyNet: A Dataset for Encouragement and Sympathy Detection. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5444–5449. Marseille, France: European Language Resources Association.

Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021. Introducing CAD: the Contextual Abuse Dataset. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *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.

Walton, D. 2005. *Fundamentals of critical argumentation*. Cambridge University Press.

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.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Yu, X.; Blanco, E.; and Hong, L. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *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.

Yu, X.; Blanco, E.; and Hong, L. 2024. Hate Cannot Drive Out Hate: Forecasting Conversation Incivility following Replies to Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 1740–1752.

Yu, X.; Zhao, A.; Blanco, E.; and Hong, L. 2023. A Fine-Grained Taxonomy of Replies to Hate Speech. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7275–7289. Singapore: Association for Computational Linguistics.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 1415–1420. Minneapolis, Minnesota: Association for Computational Linguistics.

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 Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.

Zhu, W.; and Bhat, S. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 134–149. Online: Association for Computational Linguistics.

Ziems, C.; Dwivedi-Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2023. NormBank: A Knowledge Bank of Situational Social Norms. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), 7756–7776. Toronto, Canada: Association for Computational Linguistics.

Ziems, C.; Li, M.; Zhang, A.; and Yang, D. 2022. Inducing Positive Perspectives with Text Reframing. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3682–3700. Dublin, Ireland: Association for Computational Linguistics.

Paper 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 Abstract and Introduction
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section 3-6
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Section 3-6
   (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 Ethical Statements after the checklist
   (g) Did you discuss any potential misuse of your work? Yes, see the Ethical Statements
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Ethical Statements
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we will release the code and data upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Section 6 and Appendices
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (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 Appendices
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Section 3, Section 6 and Appendices
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Section 4 and Limitations
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see Section 3-6
   (b) Did you mention the license of the assets? Yes, see the Ethical Statements
   (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 Statements
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Ethical Statements
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see Section 3 and Section 4
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, see Section 4

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

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

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

Additional Details to Identify Antisocial and Prosocial Comments

Referring to Section *Corpus Analysis*, we use an off-the-shelf RoBERTa-base model (125M parameters) from Hugging Face (Wolf et al. 2020). We show the performance of each antisocial and prosocial classifiers in Table 5. The label is *Yes* when each antisocial or prosocial behaviors is considered as observed by the classifier and *No* when it is not observed by the classifier. Note a post is identified as *Yes* in a_4 (norm violations) when the content is "*deleted*" or "*removed*", and thus we did not train a classifier.

Additional Details to Forecast Conversation Incivility Level

Details and Hyperparameters Referring to Section *Experiments*, we use an off-the-shelf RoBERTa-base model from Hugging Face (Wolf et al. 2020) to train a classifier that predict conversation incivility levels. We run the experiments on a single NVIDIA GeForce RTX 4090 GPU. It takes approximately 8 minutes to train one epoch. Table 6 shows the hyperparameters that yield the highest F1 score in predicting conversation incivility levels.

Additional Results with FLAN-T5 To minimize the variations by different models, we run experiments with FLAN-T5-base using the same experimental setting as RoBERTa-base. Table 7 shows the results. In general, FLAN-T5 achieve very similar results to RoBERTa on our task.

Experimental Details with LLMs We experiment with GPT-4o and call the API from Microsoft Azure. Figure 5 shows the zero-shot prompts and Figure 6 shows the few-shot prompts. We set the *temperature* to 0.1 and *top_p* to 0.1.

```
You will be provided with a reply to a hateful post. Will it result in Low, Medium, or
High incivility in the follow-up conversation? Answer Low, Medium, or High only. See below
all the possible labels and their description.

Label: Low
Description: low incivility in the follow-up conversation

Label: Medium
Description: medium incivility in the follow-up conversation

Label: High
Description: high incivility in the follow-up conversation

Here is the reply that needs to be classified:
Reply: ``<Reply from corpus>''
Label:
```

Figure 5: Template to generate zero-shot prompts for GPT-4o.

```
You will be provided with a reply to a hateful post. Will it result in Low, Medium, or
High incivility in the follow-up conversation? Answer Low, Medium, or High only. See below
all the possible labels and their description.

Label: Low
Description: low incivility in the follow-up conversation

Label: Medium
Description: medium incivility in the follow-up conversation

Label: High
Description: high incivility in the follow-up conversation

See below a couple of examples.

Reply: Fuck you centrist bullshit. You may as well be on the right because you are as just
as much a fucking idiot.
Label: High

Reply: Well apparently the original black panther comic from the 60s wasn't named after
the party.
Label: Medium

Reply: Yeah, because modern science says it's a bit more complicated than that| and
confusing the difference between sex and gender doesn't make him right.
Label: Low

Here is the reply that needs to be classified:
Reply: ``<Reply from corpus>''
Label:
```

Figure 6: Template to generate few-shot prompts for GPT-4o.

	Yes			No			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1
RoBERTa training with									
a_1 (offensive language)	0.86	0.78	0.82	0.94	0.96	0.95	0.92	0.92	0.92
a_2 (explicit hate speech)	0.92	0.89	0.90	0.95	0.97	0.96	0.94	0.94	0.94
a_3 (abusive language)	0.66	0.48	0.55	0.90	0.95	0.93	0.86	0.87	0.86
a_4 (norm violations)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p_1 (empathy)	0.77	0.87	0.81	0.96	0.93	0.95	0.92	0.92	0.92
p_2 (norms)	0.57	0.32	0.41	0.87	0.95	0.91	0.82	0.84	0.83
p_3 (positiveness)	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96
p_4 (politeness)	0.73	0.53	0.62	0.86	0.94	0.90	0.83	0.84	0.83

Table 5: Detailed results obtained with RoBERTa of each antisocial and prosocial behavior.

	Epochs	Batch size	Learning rate
reply	5	16	3e-5
+ pretraining	3	16	2e-5
+ blending	3	16	2e-5

Table 6: Hyperparameters used to fine-tune RoBERTa individually for each training setting.

	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	0.52	1.00	0.68	0.00	0.00	0.00	0.27	0.52	0.35
FLAN-T5 training with												
reply	0.48	0.34	0.40	0.57	0.85	0.69	0.35	0.09	0.14	0.50	0.54	0.48
+ pretraining	0.47	0.36	0.41	0.58	0.84	0.68	0.38	0.10	0.15	0.50	0.54	0.49
+ blending	0.44	0.43	0.43	0.59	0.76	0.66	0.37	0.15	0.22	0.50	0.53	0.50
hate + reply	0 .45	0.43	0.44	0.59	0.81	0.68	0.40	0.11	0.17	0.51	0.55	0.50
+ pretraining	0.42	0.56	0.48	0.62	0.70	0.66	0.39	0.16	0.23	0.52	0.53	0.51
+ blending	0.45	0.49	0.46	0.60	0.78	0.68	0.39	0.12	0.19	0.51	0.55	0.51

Table 7: Detailed results (F1 score) obtained with FLAN-T5 per label. These results complement Table 3.