Assessing the Effectiveness of Replies to Hateful Content at Stopping the Spread of Hatred

Anonymous ACL submission

Abstract

User-generated counter hate speech is a promising means to combat hate speech, but questions about effectiveness linger. Rather than identifying counter hate, this study assesses the effectiveness of replies to hateful content at stopping the spread of hatred. We estimate effectiveness based on the discourse following a reply: number of messages and how many are hateful. We argue that what matters is stopping hatred—counter hate speech that elicits more hate is counterproductive. A linguistic analysis draws insights into the language of highly, somewhat, and barely effective replies. Experimental results show that forecasting effectiveness at stopping hate is challenging. We close with a qualitative analysis shedding light into the most common errors made by the best model, including hateful content that is highly effective and polite counter hate arguments that are barely effective.

1 Introduction

The pervasive problem of online hate speech has motivated researchers to investigate methods for mitigating hatred. For example, hate speech detection has received considerable attention (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018a). Counter hate speech, which is a "direct response that counters hate speech" (Mathew et al., 2019), is a remedy to address hate speech (Richards and Calvert, 2000). Unlike content moderation, counter hate 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).

Social media platforms such as Facebook have started counter hate speech programs.[1] Recently, the NLP community has contributed several corpora with counter hate content generated on-demand by crowdworkers (Mathew et al., 2019; Qian et al., 2019) or trained operators (Chung et al., 2019).

[1] https://counterspeech.fb.com/en/

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

(Ineffective) counter hate post: *Come on man, most feminists are ok. Hate group? how can you use such a strong term?*

Hateful post (after ineffective counter hate speech): *No, it's not strong. Don't lie through your teeth. Let me know when you want to talk, c**t.*

Figure 1: An excerpt from a Reddit conversation. The second post contains counter hate speech but it is ineffective at stopping the hatred. While it may not be the cause, the counter hate post elicits additional hate. Indeed, the third post escalates the hate with respect to the original hateful post.

Researchers have used these corpora for counter hate speech detection (Mathew et al., 2019) and generation (Tekiroğlu et al., 2020; Fanton et al., 2021; Zhu and Bhat, 2021). These and other previous efforts (Section 2) make the following assumption: counter hate speech is an effective solution to address hate speech. In other words, they assume that counter hate speech can stop—or at least mitigate—hatred. While intuitive, we are not aware of strong evidence supporting this assumption. Consider the Reddit conversation in Figure 1.[2] The first post is hateful towards feminists in general. The second post is a strong counter hate argument if we ignore the follow-up conversation. As strong as it might be, however, it is ineffective: the next post escalates the hateful content further by attacking the author.

In this paper, we do not limit ourselves to focusing on counter hate speech. Rather, we analyze the effectiveness of all replies to hateful content at stopping hatred. Regardless of the content—short or long, offensive or polite, well-argued or fatally flawed from a logic standpoint—we consider a reply effective if the discourse that follows

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

is primarily not hateful. Our rationale is that what matters is preventing hateful content from spreading rather than coming up with elaborate counter hate arguments. Further, we argue that looking at genuine online discourse and assessing what comments elicit additional hateful content—even if they are well-meaning and polished counter hate arguments—is a worthwhile goal.

The complexity of discourse dynamics as well as the existence of unquantifiable outside influences (e.g.,user's mood) make analysis of effectiveness difficult from a causal perspective (Garland et al., 2022). In this study, we make no causal claims and take a more empirical approach to look at the relations between the use of language and its effectiveness in stopping the spread of online hatred. The work presented here could inspire actionable knowledge to prevent the spread of hate—a challenging goal as hate networks often have rapid rewiring and self-repairing mechanisms (Johnson et al., 2019). Our approach could further serve as a salient complement for improving current counter hate speech detection and generation systems.

We study this problem with a new dataset of Reddit conversations.[3] For the metric to measure effectiveness, following previous studies that consider hateful messages (either how many or a combined hate score) (Liu et al., 2018; Dahiya et al., 2021; Garland et al., 2022), we also take the number of non-hateful messages into account. We argue that our metric is more sound. For example, a reply to hateful content followed by five hateful posts is less effective at stopping hatred than one followed by one hateful post despite the average hate scores of the posts published after both replies are the same. Armed with the Reddit conversations and our metric to calculate effectiveness at stopping hatred, we first conduct a linguistic analysis to identify the language of the highly, somewhat, and barely effective replies to hateful posts. Additionally, we experiment with classifiers to identify whether a reply will be highly, somewhat or barely effective. Our models obtain modest results, and we present a qualitative analysis describing the most common error types. We also answer these questions:

1. Are all replies to hateful posts equally effective at stopping hatred? (they aren't);
2. Do highly, somewhat, and barely effective replies use different language? (they do);
3. Do models to assess the effectiveness of a reply benefit from having access to the hateful post in addition to the reply? (they do);
4. When differentiating between the top-k most and least effective replies, is it true that the smaller the k the easier the task? (it is).

[3]Data and code available at anonymous.link

2 Related Work

Hate speech has been an active research area in recent years (Fortuna and Nunes, 2018b). Most prior work focuses on detecting whether existing content is hateful. A few corpora have been curated for hate speech detection from diverse sources such as Twitter (Waseem and Hovy, 2016; Davidson et al., 2017), Yahoo! (Nobata et al., 2016), Fox News (Gao and Huang, 2017), Gab (Mathew et al., 2021), and Reddit (Qian et al., 2019).

There are several efforts on forecasting whether online content will result in additional hateful content. Cheng et al. (2017) build models to predict whether a moderator will flag a post for removal. Zhang et al. (2018) predict whether a comment-reply pair at the very beginning of a conversation will lead to a personal attack. Liu et al. (2018) conduct a study to forecast whether an Instagram post will receive more than n hateful comments. Dahiya et al. (2021) use a tweet and a few of the initial replies to forecast the hate score of upcoming replies. They calculate hate score by adding the probability from a toxicity classifier and the presence of hate words. These previous works propose several options to estimate hate in future content. Unlike them, we estimate the effectiveness of a reply to hateful content at stopping hatred. Estimating future hate is only a component of estimating effectiveness: non-hateful content is a good indicator of effectiveness at stopping hate.

Counter hate speech Compared with hate speech, there are fewer studies on counter hate speech. They mainly focus on counter hate speech detection (Mathew et al., 2019; Chung et al., 2019; He et al., 2021; Garland et al., 2020) and generation (Tekiroğlu et al., 2020; Fanton et al., 2021; Zhu and Bhat, 2021). There are also efforts to classify counter hate speech into fine-grained categories (Mathew et al., 2019; Chung et al., 2021).

The difficulty of building large corpora for counter hate speech is a substantial burden. He et al. (2021) use a collection of keywords relevant to COVID-19 and create a dataset with 359 tweets countering hate speech toward the Asian community. The size of their dataset, however, is small

(2k). Mathew et al. (2019) work with hateful videos towards groups and counter hate content from the video comments. Their dataset does not specify the hateful content countered in the comments, making it difficult to understand the counter hate content. Lastly, Chung et al. (2019) collect synthetic counter hate speech generated on-demand by trained operators. Compared to genuine counter hate speech written by regular people out of their own desires and motivations, synthetic counter hate speech is not as rich (e.g., *This kind of language is inappropriate and should be avoided*). Rather than defining and either detecting or generating counter hate speech, in this paper we analyze the user-generated content that stops the spread of hatred *in real online conversations*. Since we measure effectiveness at stopping hatred automatically, we bypass the burden of manual annotations.

Small-scale user studies have been conducted to investigate the effectiveness of counter hate speech by comparing the outcomes from a control group with the treatment group that that has received interventions (Munger, 2017; Hangartner et al., 2021; Bilewicz et al., 2021). The only large-scale work is by Garland et al. (2022). They work with German tweets and estimate effectiveness at stopping hatred by comparing the average hate scores of all content before and after. Unlike them, we (a) calculate effectiveness taking into account the number of all posts a reply elicits as opposed to the average hate scores, and (b) emphasize the differences in the use of language based on their effectiveness.

3 A Metric for Measuring Effectiveness at Stopping Hatred

We propose a new metric to measure how effective a reply r to hateful content is at stopping the hatred. Our metric consists of two main components: popularity and hatefulness. Popularity, denoted $P(r)$, refers to the number of comments published after r. Hatefulness, denoted $H(r)$, refers to the number of comments published after r which are hateful. Intuitively, the more hateful comments the least effective r is, but high popularity could neutralize some of the hateful comments. We formally define the effectiveness score of r as follows:

$$S(r) = \alpha P(r) - (1-\alpha)H(r) + \lambda$$

Where

$$\lambda = \begin{cases} c, & \text{if } H(r) = 0 \\ 0, & \text{if } H(r) > 0 \end{cases}$$

The parameter α determines how much weight to give to each component. We later experiment with $\alpha = 0.2$, as we believe that hateful content is more critical and several non-hateful messages are needed to neutralize hateful content. The parameter λ is a constant c (0.8 in our setting) to be added only when there is no hateful comments after r. This is to give a "reward" in the case that r does not receive any hateful comments. In other words, the greater c is, the more non-hateful follow-up comments are needed to neutralize one hateful comment. However, we do not claim that our setting is the only option or the best one. We note that absolute effectiveness scores are not as important as relative values. Here are two examples:

- If there are ten comments after r and none of them are hateful ($P(r) = 10$, $H(r) = 0$), the score is $S(r) = 2.8$.
- If there are three comments after r and two of them are hateful ($P(r) = 3$, $H(r) = 2$), the score is $S(r) = -1$.

The next section describes an application of this metric using online content from Reddit and discusses a manual validation.

4 A Corpus of Online Hate, Replies, and their Effectiveness at Stopping Hate

We choose Reddit as the starting point for our corpus. The PushShift API makes it possible to retrieve whole conversation threads seamlessly.[4] 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 of finding hateful content. Keywords, however, may introduce topic and author biases (Wiegand et al., 2019; Vidgen et al., 2021). In this study, we use community-based sampling and identify 35 subreddits (see the full list in Appendix B) that are thought to be hateful (Qian et al., 2019; Guest et al., 2021; Vidgen et al., 2021). This includes subreddits such as *r/MensRights*, *r/PurplePillDebate*, *r/ImGoingToHellForThis*, and *r/Seduction*. We retrieve a total of 1,382,596 comments from 5,325 submissions.

The next steps are to (a) identify the comments that are hateful and their replies (Section 4.1), and (b) assess how effective the replies are at stopping hatred (Section 4.2). As we shall see, the second step requires identifying hateful content in the comments following the reply.

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

4.1 Identifying Hate Comments and Their Replies

We identify hateful content in the 1,382,596 coments using pre-trained models (Liu et al., 2019) with the corpus by Qian et al. (2019) and the implementation by Pruksachatkun et al. (2020). We made this choice for several reasons. First, the corpus annotates Reddit comments as hateful or not hateful, the same domain we work with. Second, our 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). In other words, the predictions from our classifier to identify hate are reliable enough to be considered as 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 all future comments in the same thread. We found 20,286 hate comments in the 35 subreddits we work with. On average, a hate comment has 1.50 direct replies, and there are 1.94 comments published after each reply.

4.2 Assessing Effectiveness at Stopping Hatred

The metric to measure the effectiveness of a reply r at stopping hate requires us to calculate how many comments are published after r (popularity) and how many of those are hateful (hatefulness). Popularity is a simple count of the comments published after r. To increase recall, we calculate hatefulness based on the output of three classifiers. We build three models by training the same architecture as that in Section 4.1 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 a reply as hateful if any of the three classifiers predicts *hate*.

After calculating popularity and hatefulness for each reply, calculating the effectiveness score $S(r)$ is straightforward. We also use the scores to group all replies into the following categories:

- *Highly* effective if $S(r) > c$;
- *Somewhat* effective if $S(r) = c$; and
- *Barely* effective if $S(r) < c$.

We will refer to this grouping (highly, somewhat or barely) as the effectiveness level of a reply.

	Effectiveness level			
	Highly	Somewhat	Barely	All
#	8,884	15,171	6,430	30,485
%	29.1	49.7	21.2	100

Table 1: Label distribution of effectiveness levels. Almost half of replies to hateful comments are somewhat effective. There are substantially more highly effective than barely effective replies (29.1% vs. 21.2%).

Note *Barely* effective reply always has at least one hateful follow-up comments, while *Highly* effective reply usually has no hateful follow-up comments, or even it has, there are enough non-hateful comments to neutralize the few hateful ones. Our main focus is exploring the differences in the use of language between *Highly* effective and *Barely* effective replies (Section 5 and 6).

4.3 Manual Validation

We manually validate the soundness of our metric with a sample of 500 replies (250 highly effective and 250 barely effective). First, we select Reddit snippets containing the hateful comment each of the replies in the sample replies to, the reply, and all comments published after the reply. Second, we pair snippets with highly effective replies and snippets with barely effective replies (250 pairs). Third, we show the snippets to an annotator and ask her which one of the two replies is more effective at stopping hatred. The annotator agrees with the ground truth obtained with our effectiveness metric on 91.6% of the pairs, and Cohen's κ is 0.83, which is considered (almost) perfect.

5 Corpus Analysis

Table 1 presents distribution of effectiveness levels in our corpus. Almost half of the replies to a hateful comment are somewhat effective (49.7%), and 99.3% of them do not have any follow-up comment (popularity is 0). There are many highly effective replies at stopping hatred (29.1%), although a substantial amount (21.2%) are barely effective. We note that these percentages are encouraging, as the subreddits we work with are known to harvest conversations about hateful topics.

We show examples of each effectiveness level in Table 2. In the first example, the reply denounces the hate comment as sexist without getting into details or personal attacks ("[...] This statement is sexist."). This strategy is a common counter

Hate comment: *All women are feminazis. Emotional, hormonal, and anti-men. It's how they were born.* Reply r: *"all women/men are something." This statement is sexist.*	
Effectiveness score $S(r) = 5.2$, Highly effective	Popularity(r) = 22, Hatefulness(r) = 0
Hate comment: *You are a dumb motherf**ker. Please get off Reddit.* Reply r: *Stop feeding him. This is the reaction he was looking for and you're indulging him.*	
Effectiveness score $S(r) = 0.8$, Somewhat effective	Popularity(r) = 0, Hatefulness(r) = 0
Hate comment: *Lol this thread is full of internet losers like you.* Reply r: *Ha! You don't even know me little man.*	
Effectiveness score $S(r) = -16.4$, Barely effective	Popularity(r) = 33, Hatefulness(r) = 26

Table 2: Examples from our corpus and their effectiveness levels. We also include the Popularity and Hatefulness.

hate strategy (Mathew et al., 2019) and it is very successful: 22 comments follow the reply and none of them are hateful. In the second example, the reply tries to mediate the fight by persuading the author of the hate comment to stop venting ("Stop feeding him. [...] you're indulging him."). There are no comments after this reply, yielding a 0.8 effectiveness score. It is *somewhat* effective at stopping hatred—no additional hate is posted. In the third example, the reply could be considered counter hate speech as it attacks the argument in the hate comment (*full of internet losers* vs. *You don't even know me*). The reply, however, also attacks the author of the hate comment (*little man*). This reply is not effective at stopping hate: out of 33 comments that follow the reply, 26 are hateful.

Linguistic insights We perform a linguistic analysis to shed light on the language people use in the replies belonging to each effectiveness level. We split the replies into two categories (referential: Yes, or No) depending on whether they refer to the hateful content or its author. As we shall see, pretraining with this task is useful to learn models. We define it as follows:

- Quotes. Inspired by Chakrabarty et al. (2019) and Jo et al. (2020), we consider that the reply quotes the hateful comment if (a) it uses the character '>' and the text that follows overlaps with the hateful comment or (b) there are at least 4 content words in common.
- Questions. We check for question marks in the reply, as *questions* is a counter speech strategy (Chung et al., 2021).
- Negation. We check for negation cues using the list by Fancellu et al. (2016), as it is often used to dispute the hateful comment.
- Second person pronouns. We check for presence of *you* and *your*, as it is used in the replies to refer to the author of the hateful comment.

Table 7 in Appendix C presents the analysis. All factors we consider are based on counts of (a) textual features (top block) or (b) sentiment and cognition words presence. For sentiment and cognition, we use the Sentiment Analysis and Cognition Engine (SEANCE) lexicon, a tool for psychological linguistic analysis (Crossley et al., 2017). We combine a list of hate words[5] and profanity words[6] to count the profanity words.

How does language differ between highly and barely effective replies? The first pairwise comparison in Table 7 presents the answers to this question. We draw several interesting conclusions:

- When analyzing all the replies, the more tokens, nouns, verbs, negations, quotations and other textual factors indicate that the reply is barely effective. Question marks are often part of rhetorical questions and exclamations are usually not part of well-reasoned arguments but rather personal attacks.
- The textual factors are substantially different depending on whether the reply refers to the hateful comment. For example, most are no longer significant (verbs, negations, quotations, etc.) and more tokens and nouns indicate that a reply is highly effective rather than barely effective when the reply does not refer to the hateful comment.
- Regarding sentiment, negative and positive words indicate barely and highly effective replies. Similarly, profanity, and negative emotions (disgust, angry) are common in barely effective replies.

What language in the replies elicits additional comments? The second and third pairwise com-

[5] https://hatebase.org/

[6] https://github.com/RobertJGabriel/google-profanity-words-node-module/blob/master/lib/profanity.js

parisons analyze the language used in the replies that elicits comments. Recall that there is no comments after 99% of somewhat effective replies.

Despite the language of highly and barely effective replies is substantially different, the differences between somewhat effective and either highly or barely effective replies are similar: more tokens, nouns, verbs, and all other textual factors in a reply indicating that it will elicit additional comments, except exclamation marks. There are also a few interesting differences:

- More exclamation marks indicate that it will not elicit comments that are not hateful.
- More positive words indicate that it will not elicit comments that are hateful.
- Profanity, disgust and angry words are good indicators of whether the reply will elicit hateful comments (and thus be barely effective).

6 Experiments and Results

We experiment with models to solve two problems:

- Determining the effectiveness level of a reply to hateful content: highly, somewhat or barely effective (Section 6.1); and
- Differentiating between the top-$k\%$ and bottom-$k\%$ replies according to their effectiveness scores (Section 6.2).

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. The supplementary materials provide details about the hyperparameters and tuning process.

6.1 Determining Effectiveness Level

We experiment with neural classifiers built on top of RoBERTa (Liu et al., 2019). 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 (highly, somewhat, or barely effective). 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 (Qian et al., 2019; Davidson et al., 2017); (b) sentiment: negative, neutral, or positive (Rosenthal et al., 2017); (c) sarcasm: sarcasm or not sarcasm (Ghosh et al., 2020); (d) counter hate speech: hate, neutral, or counter-hate (Yu et al., 2022); (e) stance: agree, neutral, or attack (Pougué-Biyong et al., 2021); and (f) referential: yes, or no (Section 5).

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 et al., 2017; Pougué-Biyong et al., 2021; Yu et al., 2022).

6.1.1 Quantitative Results

We split the 30,485 replies in our corpus into training (70%), development (15%) and testing (15%). We present results with the testing split in Table 3. The majority baseline always predicts *somewhat*. 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: referential (reply + pretraining), counter hate speech (reply + pretraining + blending, hate comment + reply + pretraining), and stance (hate comment + reply + pretraining + blending, and + blending using reply and hate comment + reply). 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 blending without pretraining and 1.0 when pretraining and blending. The supplementary

	Highly			Somewhat			Barely			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	0.33
RoBERTa classifier with												
hate comment	0.37	0.34	0.35	0.60	0.70	0.65	0.36	0.26	0.30	0.48	0.50	0.49
reply	0.40	0.46	0.43	0.67	0.64	0.65	0.37	0.33	0.35	0.52	0.52	0.52
+ blending	0.40	0.51	0.45	0.69	0.62	0.65	0.39	0.34	0.36	0.54	0.56	0.53
+ pretraining	**0.41**	**0.59**	**0.48**	0.37	0.32	0.34	0.69	0.59	0.64	0.56	0.53	0.54
+ blending	**0.41**	**0.57**	**0.48**	0.72	0.59	0.65	0.39	0.34	0.37	0.56	0.53	0.54
hate comment + reply	0.38	0.56	0.45	0.67	0.57	0.61	0.38	0.26	0.31	0.52	0.50	0.50
+ blending†‡	0.42	0.54	0.47	**0.71**	**0.65**	**0.68**	0.41	0.32	0.36	0.56	0.55	0.55
+ pretraining	0.40	0.43	0.41	0.75	0.58	0.66	0.34	0.47	0.39	0.56	0.51	0.53
+ blending†‡	0.43	0.52	0.47	**0.74**	**0.64**	**0.68**	**0.39**	**0.41**	**0.40**	**0.58**	**0.55**	**0.56**

Table 3: Results obtained with several models. We indicate statistical significance (McNemar's test (McNemar, 1947) over the weighted average) as follows: † and ‡ indicate statistically significant ($p < 0.01$) results with respect to the model trained with the *reply* and *hate comment + reply* using neither blending nor pretraining. Training with the *hate comment + reply* coupled with pretraining with stance and blending stance yields the best results (F1: 0.56). The supplementary materials detail the results pretraining with and blending all related tasks we consider.

		Top-k%			Bottom-k%			Weighted Average		
	Size	P	R	F1	P	R	F1	P	R	F1
$k = 5$	3,657	0.79	0.70	0.74	0.56	0.68	0.61	0.71	0.69	0.69
$k = 10$	6,462	0.59	0.70	0.64	0.67	0.55	0.60	0.63	0.62	0.62
$k = 15$	10,383	0.63	0.64	0.63	0.60	0.59	0.59	0.62	0.62	0.62
$k = 20$	15,136	0.62	0.84	0.71	0.57	0.28	0.38	0.60	0.61	0.57

Table 4: Experimental results differentiating the top-k% and bottom-k% replies to hateful content according to their effectiveness scores. We present results for several values of k. The results are higher than when also identifying *somewhat effective* replies. Additionally, it is easier to differentiate the replies which have the very top and bottom of the effectiveness scores: the lower the k, the higher the weighted average.

materials present additional results.

Using only the hate comment as input is a strong baseline: it substantially outperforms the majority baseline (F1: 0.49 vs. 0.33). The reply alone yields better results (F1: 0.52); and using both the hate comment and reply without blending or pretraining is detrimental (F1: 0.50). Pretraining and blending, however, yields the best results when using the hate comment and the reply (F1: 0.56). These results lead to the following observations:

- Pretraining and blending (by themselves and combined) are more beneficial when the input is both the hate comment and the reply.
- While different systems obtain the same (or almost the same) F1 for individual labels, the system trained with the hate comment and reply, and using pretraining and blending, yields the best results overall (F1: 0.56).

6.2 Differentiating between the Top-k% and Bottom-k% replies

Although determining the effectiveness level of any reply to hateful comment is a worthwhile goal (Section 6.1), differentiating between the top-k% and bottom-k% replies according to their effectiveness scores may lead to better actionable knowledge. Indeed, the most and least effective replies are more informative than the large amount of highly and barely effective replies. Indeed the latter have a large range of effectiveness scores.

Table 4 presents the results with several k values and the best performing system from Table 3. We include in the top-k% and bottom-k% of all replies with the threshold effectiveness scores. The results show that the smaller the k, the easier it is to differentiate between the two kinds of replies. In other words, the most and least effective replies differ in language usage and the classifier is able to distinguish them. This is especially true for the top-k% most effective replies when $k = 5$.

Rhetorical question (23%)	
Hate: *Is that all you got you facist piece of sh*t?* Reply: *I don't bother with people not caring for the lives of others.*	Gold: Highly [p:1, h:0]; Pred.: Somewhat
Hateful but most effective (18%)	
Hate: *Does he create a sub for your vapid [. . .] Lol y'all are some of the dumbest mother f**kers in the world.* Reply: *You are the dumbest. Then the rest follow.*	Gold: Highly [p:6, h:0]; Pred.: Barely
Non hateful and least effective (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 other people cancer is taking it too far. Mind rule 4, please.*	Gold: Barely [p:3, h:1]; Pred.: Highly
Sarcasm or irony (15%)	
Hate: *No you retard, where is the f**king lie?* Reply: *Name calling nice argument.*	Gold: Barely [p:3, h:2]; Pred.: Highly
General knowledge (10%)	
Hate: *lol bet you thought a single thing you said in your thread wasn't retarded.* Reply: *This place is infested with incels and TD trolls.*	Gold: Barely [p:2, h:1]; Pred.: Somewhat
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, just like not all Christians are bigots.*	Gold: Somewhat [p:0, h:0]; Pred.: Barely

Table 5: Most common error types made by the best model (using as input the hate comment and the reply, and pretraining and blending). We also show the values of *popularity* (p) and *hatefulness* (h).

7 Qualitative analysis

When determining the effectiveness 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 hate comment is used to point out inappropriate content rather than expecting an answer.

The second and third most common error types (18% and 16%) are when a reply is (a) hateful but highly effective or (b) non-hateful but barely effective. Using hateful language is a common counter speech strategy (Mathew et al., 2019), and the model fails to recognize when doing so is highly effective at stopping hate. Similarly, the model struggles when countering hate politely is ineffective. When correcting misstatements, language toxicity may increase (Mosleh et al., 2021). The effectiveness may be affected by other factors such as the user identities or stances they hold, which is another reason the model struggles.

Sarcasm and irony are also common error types (15%) in our task, consistent with the task of detecting hate (Nobata et al., 2016; Qian et al., 2019). In the example, using sarcasm to point out a bad argument elicits further hate (2 out of 3 comments after the reply are hateful) and the model errs.

Errors may also occur (10%) when general knowledge is required to identify hate content that does not use offensive language. For example, 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 somewhat effective: negation does indicate barely effective replies in general (Table 7).

8 Conclusions and Future Work

Not all replies to hateful content are equally effective at stopping hatred. Indeed, replies that counter the hateful content sometimes elicit additional hate. In this paper, we work with a large dataset from Reddit and present a metric to measure effectiveness at stopping hatred. Our metric is simple and combines popularity (how many posts are published after a reply?) and hatefulness (how many are hateful?). Regardless of whether replies counter hateful content convincingly, we believe it is worthwhile to identify what kind of user-generated content is most effective at stopping hatred. While we make no causal claims which linguistic features could affect effectiveness, our analysis shows that the language of user-generated

replies differs depending on their effectiveness. For example, longer replies and those with profanity, negative, disgust or angry words are barely effective. Experimental results with transformers show that the task of forecasting effectiveness is hard to automate. We also show that pretraining and blending existing corpora yield small improvements.

Our work has several limitations. First, we identify hateful content automatically with classifiers. These classifiers obtain good results but are not perfect. As a result, some of the original hateful content we work with is actually not hateful. Second, we focus on the use of language in this work, while effectiveness may also be affected by other factors such as topics, positions, user identities, etc. We will explore them in our future research. Finally, we only consider the hateful comment and the reply in our experiments. Taking into account additional context (e.g., the full conversation thread) may be beneficial and could be applied in our future work.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Michał Bilewicz, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive behavior*, 47(3):260–266.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 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)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 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, page 1217–1230, New York, NY, USA. Association for Computing Machinery.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 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*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 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.

Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Émilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2732–2742. ACM.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 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*, pages 512–515. AAAI Press.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 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)*, pages 3226–3240, Online. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018a. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Paula Fortuna and Sérgio Nunes. 2018b. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 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*, pages 1336–1350, Online. Association for Computational Linguistics.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, 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.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, page 90–94, New York, NY, USA. Association for Computing Machinery.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020. Detecting attackable sentences in arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–23, Online. Association for Computational Linguistics.

Nicola F Johnson, R Leahy, N Johnson Restrepo, Nicolas Velasquez, Ming Zheng, P Manrique, P Devkota, and Stefan Wuchty. 2019. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 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*, pages 181–190. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 369–380. AAAI Press.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

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

Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 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.

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

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153. ACM.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte,

and Doyne Farmer. 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)*.

Rob Procter, Helena Webb, Pete Burnap, William Housley, Adam Edwards, Matthew L. Williams, and Marina Jirotka. 2019. A study of cyber hate on twitter with implications for social media governance strategies. In *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019), London, UK, October 4-5, 2019*.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 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*, pages 109–117, Online. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 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)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 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)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 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*, pages 2289–2303, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 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)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 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*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 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), pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Subreddit	Hate	Reply	Future
4Chan	513	851	1,349
antiwork	610	1,396	2,256
atheism	422	715	1,648
bakchodi	91	116	191
bindingofisaac	60	91	129
BlackPeopleTwitter	59	151	239
changemyview	835	1,127	1,890
conspiracy	1,150	1,717	3,069
DankMemes	1,384	2,136	4,106
DotA2	223	415	708
Drama	261	447	654
FemaleDatingStrategy	142	312	303
Feminism	73	136	179
GenZedong	139	198	232
HermanCainAward	1,337	1,932	3,650
justneckbeardthings	840	1,362	3,441
KotakuInAction	343	549	796
lmGoingToHellForThis	124	143	211
MensRights	1,463	1,987	4,891
MetaCanada	664	828	1,546
modernwarfare	209	316	694
NoFap	60	66	75
playrust	108	142	308
PurplePillDebate	1,194	1,599	4,100
PussyPass	656	791	1,432
PussyPassDenied	3,258	4,635	9,060
Seduction	174	239	364
ShitPoliticsSays	630	819	2,283
ShitRedditSays	271	353	491
Sino	40	87	43
SubredditDrama	1,411	2,204	4,683
TrueReddit	319	440	1,096
TumblrInAction	138	184	359
TwoXChromosomes	197	637	562
worldnews	888	1,364	2,318
Total	20,286	30,485	59,356

Table 6: Number of hate comments, replies, and comments after the replies per subreddit.

A Ethical Considerations

We use the PushShift API to collect data from Reddit.[7] The collection process is consistent with Reddit's Terms of Service. We access our data through the data dumps on Google's BigQuery using Python.[8]

Reddit can be considered a public space for discussion which differs from a private messaging service (Vidgen et al., 2021). Users consent to have their data made available to third parties including academics when they sign up to Reddit. Existing 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.[9]

The annotator was warned of the potential hateful content before working on our task. We provide annotator with access to supporting services throughout the task. Annotator was compensated with 8 US$ per hour.

B Data

The number of hate comments (Hate), replies (Reply) and comments after the replies (Future) in each subreddits are detailed in Table 6. In total, there are 20,286 hate comments, 30,485 replies, and 59,356 comments after the replies from the 35 subreddits.

C Linguistic Analysis

Table 7 presents the details of the linguistic analysis discussed in Section 5.

D Detailed Results

Table 8 presents detailed results complementing Table 3 in the paper. We provide Precision, Recall and weighted F1-score using each related task for pretraining and blending.

E Hyperparamters and Finetuning Process

Our dataset was pre-processed by removing URLs, removing symbols, removing any additional spaces, and at the end, converting all words to lower-case. The neural model takes about an hour on average to train on a single NVIDIA TITAN Xp. We use the implementation by Pruksachatkun et al. (2020) and fine-tune the RoBERTa (base architecture; 12 layers) (Liu et al., 2019) model for each of the four training settings. For each setting, we set the hyperparameters to be the same when the input is the hateful comment, the reply, or both (Table 9).

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

[8] https://pushshift.io/using-bigquery-with-reddit-data/

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

	Highly vs. Barely			Highly vs. Somewhat			Barely vs. Somewhat		
	All	Referential?		All	Referential?		All	Referential?	
		Yes	No		Yes	No		Yes	No
Textual factors									
Tokens	↓↓↓	↓↓↓	↑↑	↑↑↑	↑↑	↑↑↑	↑↑↑	↑↑↑	
Nouns	↓↓↓	↓↓↓	↑↑	↑↑↑			↑↑↑	↑↑↑	↑↑↑
Verbs	↓↓↓	↓↓↓		↑↑↑		↑↑↑	↑↑↑	↑↑↑	↑↑↑
Negations	↓↓↓	↓↓↓		↑↑↑	↑↑↑		↑↑↑	↑↑↑	
Quotations	↓↓↓	↓↓		↑		↑	↑↑↑	↑↑	
Exclamations marks	↓	↓		↓↓↓	↓↓↓	↓			
Questions marks	↓↓			↑↑↑	↑↑↑		↑↑↑	↑↑↑	
1st person pronouns	↓↓	↓↓		↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑
2nd person pronouns	↓↓↓	↓↓↓	↓↓↓	↑↑↑	↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑
Sentiment and Cognition									
Negative words	↓↓↓	↓↓↓			↓↓↓		↑		↑↑↑
Positive words	↑↑↑	↑↑↑			↑		↓↓↓		↓↓↓
Profanity words	↓↓↓	↓↓↓	↓↓	↓↓	↓↓↓	↓↓↓	↑↑↑		
Overstated words		↓		↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↑↑↑
Disgust words	↓↓↓	↓↓↓			↓		↑	↑	
Angry words	↓↓↓	↑	↓↓↓				↑↑↑	↑	↑↑↑

Table 7: Linguistic analysis comparing the replies to hateful content that are highly, somewhat and barely effective at stopping hatred. We provide results for all replies in each effectiveness level and depending on whether they refer to the hateful comment. Number of arrows indicate 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 level (up) or the second (down) in each pairwise comparison. The few tests that do not pass the Bonferroni correction are underlined.

	Highly			Somewhat			Barely			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	0.33
reply												
+ pretraining with												
Hate_twitter	0.40	0.59	0.47	0.75	0.57	0.65	0.34	0.30	0.32	0.56	0.52	0.53
Hate_reddit	0.40	0.58	0.47	0.74	0.57	0.64	0.36	0.32	0.34	0.56	0.52	0.53
Sentiment	0.38	0.55	0.45	0.73	0.57	0.64	0.35	0.31	0.33	0.55	0.51	0.52
Sarcasm	0.38	0.55	0.45	0.69	0.56	0.62	0.37	0.30	0.33	0.53	0.50	0.51
Counter	0.39	0.52	0.45	0.74	0.57	0.64	0.32	0.35	0.33	0.55	0.51	0.52
Stance	0.40	0.54	0.46	0.75	0.56	0.64	0.33	0.36	0.35	0.56	0.51	0.52
Referential	**0.41**	**0.59**	**0.48**	0.37	0.32	0.34	0.69	0.59	0.64	0.56	0.53	0.54
+ blending with												
Sentiment	0.39	0.49	0.44	0.72	0.59	0.65	0.38	0.39	0.38	0.55	0.52	0.53
Counter	0.40	0.56	0.47	0.70	0.61	0.65	0.40	0.31	0.35	0.55	0.53	0.53
Stance	0.40	0.51	0.45	0.69	0.62	0.65	0.39	0.34	0.36	0.54	0.56	0.53
+ pretraining + blending												
Sentiment	0.40	0.46	0.43	0.66	0.66	0.66	0.37	0.30	0.33	0.52	0.52	0.52
Counter	**0.41**	**0.57**	**0.48**	0.72	0.59	0.65	0.39	0.34	0.37	0.56	0.53	0.54
Stance	0.39	0.50	0.44	0.69	0.62	0.65	0.38	0.33	0.36	0.54	0.52	0.53
hate comment + reply												
+ pretraining with												
Hate_twitter	0.40	0.45	0.42	0.68	0.62	0.65	0.37	0.39	0.38	0.53	0.52	0.53
Hate_reddit	0.40	0.58	0.47	0.74	0.57	0.64	0.36	0.32	0.34	0.56	0.52	0.53
Sentiment	0.39	0.34	0.36	0.58	0.75	0.65	0.42	0.21	0.28	0.49	0.51	0.49
Sarcasm	0.38	0.55	0.45	0.69	0.56	0.62	0.37	0.3	0.33	0.53	0.50	0.51
Counter	0.40	0.43	0.41	0.75	0.58	0.66	0.34	0.47	0.39	0.56	0.51	0.53
Stance	0.38	0.49	0.43	0.67	0.55	0.60	0.37	0.38	0.37	0.52	0.50	0.50
Referential	0.39	0.51	0.44	0.37	0.34	0.35	0.69	0.59	0.64	0.53	0.51	0.52
+ blending with												
Sentiment	0.38	0.42	0.40	0.67	0.63	0.65	0.38	0.38	0.38	0.52	0.52	0.52
Counter	0.38	0.41	0.40	0.64	0.66	0.65	0.38	0.31	0.35	0.51	0.51	0.51
Stance	0.42	0.54	0.47	**0.71**	**0.65**	**0.68**	0.41	0.32	0.36	0.56	0.55	0.55
+ pretraining + blending												
Sentiment	0.40	0.45	0.42	0.68	0.60	0.64	0.37	0.41	0.39	0.53	0.52	0.52
Counter	0.40	0.47	0.43	0.71	0.56	0.63	**0.36**	**0.45**	**0.40**	0.55	0.51	0.52
Stance	0.43	0.52	0.47	**0.74**	**0.64**	**0.68**	**0.39**	**0.41**	**0.40**	**0.58**	**0.55**	**0.56**

Table 8: Detailed results (P, R, and F) predicting whether the reply is Highly, Somewhat, and Barely when the input is only the reply or the hate comment + reply. These results are using RoBERTa and pretrained with or blending each each related task. This table complements Table 3 in the paper.

	Epochs	Batch size	Learning rate	Dropout	Patience
reply	5	8	1e-5	0.5	10
+ blending	5	8	1e-5	0.5	10
+ pretraining	5	8	1e-5	0.5	10
+ blending	4	8	1e-5	0.5	10

Table 9: Hyperparameters used to fine-tune RoBERTa individually for each training setting. We accept default settings for the other hyperparameters as defined in the implementation by Pruksachatkun et al. (2020).