Echoes of Discord: Forecasting Hater Reactions to Counterspeech

Xiaoying Song[1] Sharon Lisseth Perez[1] Xinchen Yu[2] Eduardo Blanco[2] Lingzi Hong[1]
[1]College of Information, University of North Texas
[2]Department of Computer Science, University of Arizona
{xiaoyingsong, sharonperez}@my.unt.edu
{xinchenyu, eduardoblanco}@arizona.edu lingzi.hong@unt.edu

Abstract

Hate speech (HS) erodes the inclusiveness of online users and propagates negativity and division. Counterspeech has been recognized as a way to mitigate HS. While some research has investigated the impact of user-generated counterspeech on social media platforms, few have examined and modeled haters' reactions toward counterspeech, despite the immediate alteration of haters' attitudes being an important aspect of counterspeech. This study fills the gap by analyzing the impact of counterspeech from the hater's perspective, focusing on whether the counterspeech leads the hater to reenter the conversation and if the reentry is hateful. We compile the **Re**ddit **Echo**es of Hate dataset (**ReEco**), which consists of triple-turn conversations featuring haters' reactions, to assess the impact of counterspeech. To predict haters' behaviors, we employ two strategies: a two-stage reaction predictor and a three-way classifier. The linguistic analysis sheds insights on the language of counterspeech to hate eliciting different haters' reactions. Experimental results demonstrate that the 3-way classification model outperforms the two-stage reaction predictor, which first predicts reentry and then determines the reentry type. We conclude the study with an assessment showing the most common error causes in the best-performing model. The dataset used in this study is publicly available for further research [1].

Trigger warning: Read with caution. Examples in this paper may present toxic languages

1 Introduction

Hate speech (HS) online causes increased prejudice and discrimination, fostering an environment of hostility and social division (Chetty and Alathur, 2018). This motivates researchers to explore methods to mitigate the negative impact of HS, including

A	So you're so big of a pu*** that you know something is wrong but the smell of a fast food restaurant makes it impossible for you to stop? Jesus Christ you're a little b***.
B	People like you are why vegans get bad rep, it doesn't help to chastise the very people you're trying to convert. I'd stay away from the evangelizing aspect of veganism and let others do that work because you're clearly really not good at it, in fact people like you are detrimental to the cause.
A	I don't care. Stop being a b*** and doing something you know is wrong. I don't have to coddle an abuser

Table 1: An example of hater reentry from a Reddit conversation. User A posted a hate comment. User B replied to the hate speech, which induced A to reenter the conversation with a hateful post.

the automatic detection of HS (Abro et al., 2020; Duwairi et al., 2021) and generation of counterspeech (Saha et al., 2022; Tekiroglu et al., 2022).

Directly addressing HS through counterspeech is regarded as an effective strategy to curb the spread of hatred and promote constructive dialogue (Bonaldi et al., 2022). Previous research has employed crowdsourced workers (Qian et al., 2019) or NGO experts (Chung et al., 2021) to craft counterspeech datasets, which are used to develop models for automatic counterspeech generation. However, the actual effect of synthetic counterspeech on social media platforms is unknown.

Indeed, counterspeech may generate unfavorable conversation outcomes. Table 1 shows a slightly altered dialogue extracted from Reddit. The initial post by User A attacked someone with hateful words. In response, User B attempted to alter User A's perspective by stating, "It doesn't help to..." However, this reply angered User A, resulting in User A's reentry into the conversation with more hateful content. While the intent was to counter hatred, the outcome was counterproductive—rather than de-escalating the situation, the response fueled further hostility, intensifying the conflict.

Several studies have proposed to assess the effectiveness of counterspeech by measuring conversation incivility (He et al., 2021; Baider, 2023; Yu et al., 2023). This measurement considers the
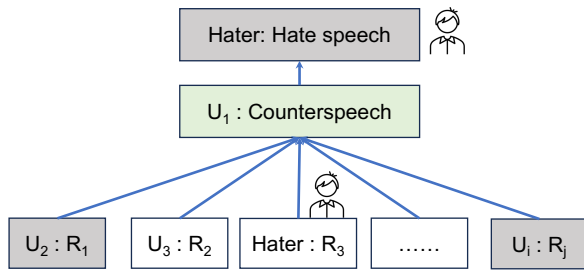
---

[1]https://github.com/oliveeeee25/counterspeech_effectiveness_hater_reentry

4907

Figure 1: Hater's non-hateful reentry as a conversation outcome. A Reddit user (hater) posts HS. Another user, U_1, replies with counterspeech. This is followed by subsequent replies (R_1, R_2, ...R_j). The counterspeech prompts the hater to reenter the conversation with a non-hateful post. Grey boxes represent HS.

reactions of all participants including bystanders.

Changing the attitudes and behaviors of haters is an important aspect of counterspeech (Stroebe, 2008). However, few studies have investigated the impact of counterspeech from the perspective of hate perpetrators. Effective counterspeech can induce cognitive dissonance in haters by challenging their beliefs or assumptions, leading them to reconsider and potentially change behavior (Dillard and Harmon-Jones, 2002). Counterspeech may make perpetrators aware of moral implications and the harmful impact of their actions, potentially re-reengaging their sense of empathy and leading to behavior change (Wachs et al., 2022). On the other hand, hate perpetrators may perceive counterspeech as coercive, which triggers their stronger attachment to original beliefs and more resistance to change (Acheme et al., 2024). It is important to identify effective counterspeech to hate perpetrators: the ones that lead to immediate moderation of the hater's rhetoric (Baider, 2023; Leonhard et al., 2018), or at least would not incite the hater to be more hateful (Shugars and Beauchamp, 2019).

In this study, we investigate haters' reactions following counterspeech to HS in real conversations on a social media platform, Reddit. Our research addresses the following questions:

- Are there linguistic differences in counterspeech that lead to different hater reactions?
- Which language models can more accurately predict hater reactions?

Examining haters' reactions can reveal which characteristics of counterspeech are more likely to encourage constructive behavior. These insights can guide users to address HS in a way that minimizes the risk of provoking further negativity. By developing models that predict haters' reactions, we can assess counterspeech, whether manually crafted by crowdsourced workers or generated by algorithms, and identify responses that are most likely to elicit desired hater reactions. The curated genuine conversation data can improve the training of large language models (LLMs) to generate counterspeech that effectively manages hater reactions.

To answer these questions, we build the **Re**ddit **Echo**es of Hate corpus (ReEco), which includes conversation pairs (HS and counterspeech) labeled by hater reactions: no reentry, hateful reentry, and non-hateful reentry. We perform linguistic analysis to identify variations in counterspeech that elicit different hater reactions. We experiment with language models and adopt two strategies for predicting haters' reactions: **(i) a two-stage reaction predictor** that is composed of two predictive models: identifying hate reentry (yes/no) followed by identifying reentry type (hateful or non-hateful). This method divides the prediction task into two stages, allowing each model to focus on a simpler, more defined problem. Considering that errors in the first model can affect the second model's input and lead to compounded inaccuracies (Lambert et al., 2022), we develop **(ii) a 3-way classifier** that predicts one of the three outcomes (no reentry, hateful reentry, or non-hateful reentry). Experiments show the 3-way classification achieves the highest prediction accuracy, while large language models (LLMs) are not superior in predicting haters' reactions compared to BERT models.

2 Related Work

Counterspeech to HS Counterspeech refers to a direct response to address HS (Mathew et al., 2019). Unlike approaches that block HS or haters, counterspeech fosters dialogues and promotes understanding without suppressing freedom of expression (Schieb and Preuss, 2016), which can effectively de-escalate tension and reduce hostility (Hangartner et al., 2021). Previous studies have curated counterspeech datasets by employing humans or NGO experts to generate synthetic counterspeech (Qian et al., 2019; Chung et al., 2019). Manual crafting of counterspeech is time-consuming and limits the capability of scaling. Models have been developed for generating counterspeech (Bonaldi et al., 2022; Chung et al., 2020, 2021; Zhu and Bhat, 2021). Few of these studies have considered the effectiveness of counterspeech from the perspective of conversation outcomes. We bridge this gap by examining user-generated coun-

Community	# Hate Speech	# Counterspeech with Follow-up	Outcomes		
			No Reentry (%)	Hateful Reentry (%)	Non-hateful Reentry (%)
Discussion	6,862	1,024	402 (39%)	148 (15%)	474 (46%)
Identity	7,779	2,247	721 (32%)	468 (21%)	1,058 (47%)
Media-sharing	2,549	1,417	435 (31%)	309 (22%)	673 (47%)
Meme	2,399	528	143 (27%)	145 (27%)	240 (46%)
Hobby	2,441	507	179 (35%)	98 (19%)	230 (46%)
All	22,030	5,723	1,880 (33%)	1,168 (20%)	2,675 (47%)

Table 2: Analysis of our corpus, ReEco.

terspeech and its impact on haters' reactions.

Effectiveness of Counterspeech Assessing the effectiveness of counterspeech is crucial for crafting appropriate counterspeech that mitigates the harms of HS (Cepollaro et al., 2023). Baider (2023) uses the number of answers triggered by the comment and the tone of the answers to determine the effectiveness of counterspeech. Yu et al. (2023) assess the effectiveness based on conversation incivility, which considers the number of hateful and non-hateful comments, along with the unique authors participating. Most of these works evaluate the effectiveness based on bystanders of the conversation. Reynolds and Tuck (2016) propose using qualitative analysis to assess the effect of counterspeech based on the hate perpetrators. This study is the first to model haters' reactions with genuine conversation data from a social media platform.

Conversation Outcome Forecasting Methods have been developed to predict the future development of a conversation (Bao et al., 2021). Previous works predict conversational outcomes from different aspects, such as users' reentry (Wang et al., 2021; Zeng et al., 2019), the popularity of a conversation (Rizos et al., 2016; Risch and Krestel, 2020; Backstrom et al., 2013), conversation killers (Jiao et al., 2018), and derailment of conversations (Chang and Danescu-Niculescu-Mizil, 2019). In particular, Backstrom et al. (2013) has used users' network relationships, and Zeng et al. (2019) has used users' historical behaviors to predict reentry behaviors. However, these studies do not provide direct insights into how counterspeech influences subsequent conversation outcomes, which are crucial for developing counterspeech generation strategies and models. Additionally, few studies have modeled haters' reactions following counterspeech.

3 A New HS/counterspeech Corpus - ReEco

Data Collection and Process We use the PushShift API to collect complete conversation threads containing HS from Reddit.[2] We employ community-based sampling, selecting 42 subreddits that are identified to exhibit a higher prevalence of HS by Qian et al. (2019), Guest et al. (2021), and Vidgen et al. (2019), including r/MensRights, r/PurplePillDebate, and r/Seduction, etc (see Appendix E for the complete list). We collect a total of 1,410,361 comments ranging from February 2, 2019, to November 26, 2021.

Detect HS HS refers to expressions in which the author deliberately targets an individual or group with the intent to vilify, humiliate, or incite hatred (Yu et al., 2022). We fine-tune RoBERTa models (Liu et al., 2019) with the HS datasets by Vidgen et al. (2021), Qian et al. (2019), and Davidson et al. (2017), and build three HS classifiers. A comment is labeled as hateful only if all three classifiers consistently predict it as such. This is to minimize the risk of any classifiers being wrong. The prediction-assigned labels are compared with human annotations to validate reliability. Two researchers manually annotated 200 samples guided by the HS definition and examples (annotation details in Appendix D). The agreement rate is 93%. The Kappa coefficient between the predicted results and manual labeling is 0.76, indicating a substantial level of agreement (Viera et al., 2005). The manual evaluation verifies the accuracy of the predictions. While the classifiers are not flawless, most of the conversations identified are related to HS, ensuring the generation of reliable results.

Identify Counterspeech Counterspeech are responses to HS which are crafted to contradict and challenge hateful remarks (Chung et al., 2023). For comments predicted to be HS, we obtain their subsequent two-level replies. There are 40,162 HS

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

	Meme	Hobby	Identity	Discussion	Media-sharing	All
Textual Factors						
2nd Person Pronouns	↑	↑	↑	↑	↑	↑
Uncertainty	↑	↑	↑	↑	↑	↑
Abstract	↑	↓	↑	↑	↑	↑
Emotional Factors						
Enlightenment	↑	↑	↑	↑	↑	↑
Negative	↑	↓	↑	↑	↑	↑
Fear	↑	↓	↑	↑	↑	↑
Positive	↓	↓	↓	↓	↓	↓
Polarity	↑	↓	↑	↑	↑	↑
Valence	↑	↑	↑	↑	↑	↑

Table 3: Linguistic analysis comparing the counterspeech that triggers hater reentry and no reentry in different communities. The up arrow indicates higher values in counterspeech with hater reentry. Tests that do not pass the Bonferroni correction are underlined.

	Meme	Hobby	Identity	Discussion	Media-sharing	All
Textual Factors						
Causation	↑	↑	↑	↑	↑	↑
Format	↑	↑	↑	↑	↑	↑
Social-related Factors						
Respect	↓	↓	↓	↓	↓	↓
Power	↓	↓	↓	↓	↓	↓
Worship	↓	↓	↓	↓	↓	↓
Forgiveness	↓	↓	↓	↓	↓	↓
Emotional Factors						
Longing	↓	↓	↓	↓	↓	↓
Exclamation	↑	↑	↑	↑	↑	↑
Aggression	↑	↑	↑	↑	↑	↑

Table 4: Linguistic analysis comparing the counterspeech that triggers hateful/non-hateful reentry in different communities. The up arrow indicates higher values in counterspeech with hateful reentry. Tests that do not pass the Bonferroni correction are underlined.

and 76,009 replies to these HS. Of these pairs, only 22,030 HS and their replies have follow-up replies, which are used to analyze hater reactions. We exclude pairs without follow-up replies as these pairs are located at the end of the dialogue tree, which could happen when there is topic exhaustion after thorough discussions or external interruptions (Jiao et al., 2018). By excluding these pairs, we can largely avoid the impact of these factors and focus on haters' 'no reentry' when the conversations are actively evolving.

Not all replies to HS qualify as counterspeech. We developed three classifiers, fine-tuned on three distinct counterspeech datasets (Albanyan et al., 2023; Yu et al., 2022; Vidgen et al., 2021), to identify counterspeech within those replies. We consider a reply to be counterspeech if the three classifiers indicate so, which results in 5,723 (HS, counterspeech) pairs from Reddit. We manually annotated 200 samples to validate the results (annotation details in Appendix D). The agreement rate is 95%. The Kappa coefficient between the predicted results and manual labeling is 0.79, indicating the result is reliable (Viera et al., 2005). We examine whether the user who posts the initial HS (i.e., hater) shows up in the follow-up conversation to counterspeech, categorizing the conversation as with hater 'reentry' or 'no reentry.' Based on the prediction of HS classifiers, the reentry comments are further labeled as 'hateful' or 'non-hateful'. Figure 1 illustrates the outcome of a conversation that involves a hater's reentry with a non-hateful reply to counterspeech.

Description of ReEco The final corpus, ReEco, consists of 5,723 (HS, counterspeech) pairs. To investigate whether the hater reentry behavior varies across communities, we group the 42 subreddits into 5 categories (see Appendix E), i.e., *Discussion*, *Identity*, *Media-sharing*, *Meme*, and *Hobby* (Weld et al., 2022), and analyze the distribution of hater reentry labels in Table 2. There are more HS in *Discussion* and *Identity*. Different communities exhibit similar proportions of *non-hateful reentry* ($\approx 46\%$), *hateful reentry* ($\approx 21\%$), and *no reentry* ($\approx 33\%$). However, there are notable variations. The *Meme* community shows a slightly higher proportion of *hateful reentries* (27%) and a lower proportion of *no reentry* (27%), whereas the *Discussion* community has a higher proportion of *no reentries* (39%) and a lower proportion of *hateful reentries* (15%).

4 Corpus Analysis

We conduct two linguistic analyses, the first one analyzes differences in counterspeech that elicit hater to reenter or not. The second one takes the counterspeech with hater reentry and investigates the differences that elicit hateful or non-hateful reentry. These analyses are carried out across different communities, as topics and sentiments may vary in communities. We use SEANCE (Crossley et al., 2017) for sentiment analysis and spaCy (Honnibal and Montani, 2017) for entity recognition. We further employ the Wilcoxon rank-sum test to perform statistical comparisons. The Bonferroni correction is applied to highlight the most significant linguistic features (Weisstein, 2004). (See details of linguistic features in Appendix C)

Table 3 details linguistic findings in counterspeech that trigger haters' reentry. Counterspeech that includes elements of enlightenment, and negative and fear emotions is more likely to trigger hater reentry, which is consistent across most com-

Model	Reentry			No Reentry			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1
Baseline	0.69	1.00	0.82	0.00	0.00	0.00	0.48	0.69	0.57
BERT									
HS	0.77	0.69	0.73	0.72	0.80	0.75	0.74	0.74	0.74
Counterspeech	0.87	0.71	0.78	0.75	0.89	0.82	0.81	0.80	0.80
Pair	0.89	0.69	0.78	0.75	0.91	0.82	0.82	0.80	0.80
BERT-MTL									
HS	0.79	0.75	0.77	0.76	0.80	0.78	0.78	0.78	0.78
Counterspeech	0.84	0.78	0.81	0.79	0.85	0.82	0.81	0.81	0.81
Pair	0.86	0.79	**0.82**	0.81	0.87	**0.84**	0.83	0.83	**0.83**
LLaMA 3 Zero-shot									
HS	0.39	0.83	0.53	0.55	0.14	0.22	0.49	0.42	0.35
Counterspeech	0.60	0.51	0.55	0.40	0.49	0.44	0.50	0.50	0.49
Pair	0.41	0.60	0.49	0.62	0.43	0.50	0.53	0.50	0.50
LLaMA 3 Finetuned									
HS	0.74	0.83	0.78	0.41	0.29	0.33	0.64	0.67	0.65
Counterspeech	0.73	0.83	0.78	0.40	0.27	0.33	0.64	0.67	0.65
Pair	0.75	0.85	0.80	0.46	0.30	0.36	0.66	0.69	0.67

Table 5: Performance comparison of different models and inputs on reentry prediction.

munities, except Meme. Counterspeech with hater reentry (except Hobby) has significantly more polarity and valence words, indicators of emotionally charged and positioned language. This phenomenon can be explained by the emotional contagion theory (Hancock et al., 2008) and the affective intelligence theory (Marcus et al., 2011). Emotions in counterspeech can lead to haters feeling and displaying similar emotions. These emotions influence how the hater processes information and engages in the conversation (Choi and Lee, 2021).

We observe some new findings that have been seldom mentioned or differ. Second-person pronounces, uncertainty (words denoting feelings of uncertainty, e.g., about, almost), and abstract words (e.g., words reflecting a tendency to use abstract vocabulary, e.g., ability, advantage) show higher prevalence in counterspeech that causes hater reentry except in Hobby. Unlike the findings of (Lubis et al., 2019), positive emotions in counterspeech tend to inhibit further engagement from haters (i.e., significantly more in no reentry). This suggests counterspeech has distinctive characteristics compared to general conversational engagement.

These observations imply that the emotional tone of counterspeech can influence haters' reactions. Ghandeharioun et al. (2019) have proposed designing emotionally aware agents to promote behavior change in wellness applications. Our analysis suggests this idea may also have value in altering haters' behaviors.

Table 4 presents the linguistic differences between counterspeech that causes different types of reentry. Significant differences are summarized into three aspects, including textual, social-related, and emotional factors. Textual factors, such as causation and format, occur more frequently in instances of hateful reentry and this trend is consistent across all communities. In social-related factors, counterspeech that prompts non-hateful reentry tends to include more elements of respect, power, worship, or forgiveness. These insights provide clues to encourage positive interactions and facilitate constructive dialogue. It is hypothesized that haters modify their behavior in such cases due to cognitive dissonance, as continuing to engage in hateful actions while receiving respect (Dillard and Harmon-Jones, 2002) and mutual understanding (Clark et al., 2019) may create psychological discomfort. Further research is needed to explore the underlying mechanisms behind how these signals prompt behavioral change in haters.

Regarding emotional factors, counterspeech with hateful reentry conveys elements like exclamation and aggression, while counterspeech with non-hateful reentry contains more longing emotion. It implies counterspeech with aggression potentially incites further hostility while expressions of longing for better understanding foster a less negative re-engagement. This further verifies that emotion in counterspeech plays a crucial role in shaping haters' reactions (Hancock et al., 2008).

	Hateful Reentry			Non-Hateful Reentry			Weighted Average		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	0.00	0.00	0.00	0.70	1.00	0.82	0.48	0.70	0.57
BERT									
HS	0.67	0.83	0.74	0.77	0.59	0.67	0.72	0.71	0.71
Counterspeech	0.68	0.91	0.78	0.86	0.57	0.69	0.77	0.74	0.73
Pair	0.74	0.89	0.81	0.87	0.68	0.76	0.80	0.79	0.78
BERT-MTL									
HS	0.77	0.86	0.81	0.84	0.75	0.79	0.81	0.80	0.80
Counterspeech	0.76	0.87	0.81	0.85	0.73	0.78	0.80	0.80	0.80
Pair	0.80	0.84	**0.82**	0.84	0.79	**0.81**	0.82	0.82	**0.82**
LLaMA 3 Zero-shot									
HS	0.59	0.86	0.70	0.42	0.14	0.21	0.52	0.57	0.50
Counterspeech	0.60	0.65	0.62	0.44	0.38	0.41	0.53	0.54	0.53
Pair	0.60	0.88	0.71	0.45	0.14	0.22	0.54	0.58	0.51
LLaMA 3 Finetuned									
HS	0.32	0.22	0.26	0.70	0.79	0.74	0.58	0.61	0.59
Counterspeech	0.36	0.27	0.31	0.71	0.80	0.75	0.61	0.63	0.62
Pair	0.40	0.16	0.23	0.71	0.90	0.79	0.61	0.67	0.62

Table 6: Performance comparison of different models and inputs on reentry type prediction.

5 Experiments and Results

We experiment with two strategies to predict haters' reactions following counterspeech. The first **Two-stage Reaction Predictor** includes two consecutive tasks: (i) predicting whether a hater reenters the conversation, and (ii) predicting whether the reentry is hateful or not, if reentry occurs. The second **3-way Response Classifier** predicts the hater's reaction as one of the three results: no reentry, hateful reentry, and non-hateful reentry.

5.1 Experiment Implementation

We split ReEco into training (80%) and testing (20%), and report model performance on the test data. For each prediction task, we experiment with the following three types of input to identify which information can best model haters' reactions.

HS Liu et al. (2018) found that hateful comments can predict future engagement in conversations. We adopt this idea and experiment with predicting haters' reactions based on the HS.

Counterspeech How people respond to HS may elicit varying responses by the hater (Shugars and Beauchamp, 2019). We explore how effectively counterspeech can predict haters' reactions.

Pair We hypothesize that the interaction between HS and counterspeech influences the hater's subsequent behavior. To model this interaction, we concatenate the HS and counterspeech using the [SEP] special token as input to BERT-large-uncased. For the LLM setup, the conversation pair is formed by concatenating the text in the query.

We experiment with the following models for prediction tasks. This is to validate the robustness of our findings, identify the most predictive information, and determine the optimal language models for accurate prediction.

BERT The model's neural architecture is built on a BERT-large-uncased base, followed by a fully connected layer with 1,024 neurons and tanh activation. This is then connected to a final fully connected layer with 3 neurons and softmax activation.

Multi-task Learning (MTL) Models MTL allows a model to learn from multiple related tasks simultaneously. By integrating both shared and context-specific representations, the model could potentially achieve enhanced performance (Zhang and Yang, 2018). We utilized a deep neural network architecture with a shared base, BERT-large-uncased, followed by task-specific layers for each classification output. The model is trained on three pertinent tasks: classification of toxicity,[3] detection of HS and counterspeech (He et al., 2021), and identification of personal attacks (Zhang et al., 2018).

Llama 3 Zero-shot Large language models (LLMs), such as Llama 3, have been trained on large corpora and demonstrated capabilities in understanding human language. We experiment with Llama 3 in dialogue to test whether the zero-shot learning model can accurately predict haters' reactions. To constrain the format of the generated response, we adopt Guidance, a programming

---

[3]Available publicly at https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

4912

	Hateful Reentry			Non-Hateful Reentry			No Reentry			Weighted Average		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	0.00	0.00	0.00	0.48	1.00	0.65	0.00	0.00	0.00	0.23	0.48	0.31
BERT												
HS	0.73	0.85	0.79	0.66	0.53	0.59	0.74	0.77	0.76	0.71	0.72	0.71
Counterspeech	0.81	0.80	0.80	0.66	0.52	0.58	0.67	0.73	0.71	0.71	0.71	0.71
Pair	0.73	0.89	0.80	0.70	0.55	0.61	0.80	0.80	**0.80**	0.74	0.75	0.74
BERT-MTL												
HS	0.81	0.83	0.82	0.68	0.56	0.62	0.71	0.82	0.76	0.74	0.71	0.73
Counterspeech	0.84	0.80	0.82	0.67	0.63	0.65	0.72	0.79	0.76	0.74	0.74	0.74
Pair	0.81	0.86	**0.84**	0.69	0.67	**0.68**	0.78	0.78	0.79	0.77	0.77	**0.77**
LLaMA 3 Zero-shot												
HS	0.44	0.10	0.16	0.28	0.44	0.35	0.39	0.57	0.46	0.38	0.35	0.31
Counterspeech	0.44	0.15	0.23	0.29	0.42	0.34	0.33	0.46	0.39	0.36	0.33	0.31
Pair	0.40	0.68	0.51	0.26	0.18	0.21	0.35	0.17	0.23	0.35	0.37	0.33
LLaMA 3 Finetuned												
HS	0.33	0.15	0.20	0.50	0.63	0.56	0.36	0.35	0.36	0.42	0.44	0.42
Counterspeech	0.35	0.17	0.23	0.52	0.66	0.58	0.42	0.37	0.39	0.45	0.47	0.45
Pair	0.37	0.17	0.24	0.52	0.67	0.59	0.42	0.39	0.40	0.46	0.48	0.46
Two-stage (BERT-MTL)												
Pair	0.13	0.27	0.17	0.73	0.75	0.74	1.00	0.26	0.42	0.70	0.51	0.53

Table 7: Performance comparison of different models and inputs for 3-way reentry prediction.

paradigm that enables constrained generation with regex,[4] on the Llama-3.1-8B-Instruct model for inferences.

The query for the reentry prediction is formulated as:

```
Here is a counterspeech to a hate
comment: <example>.
Will the hater come back to join the
conversation? Answer 'Yes' or 'No'.
```

The query for reentry type prediction is:

```
Here is a counterspeech to a hate
comment: <example>.
Assuming the hater comes back to join
the conversation, will the engagement
be hateful? Answer 'Yes' or 'No'
```

The query for the 3-way response is:

```
Here is a counterspeech to a hate
comment:<example>.
What will be the hater's response?
Answer 'Reentry with a non-hateful
comment', 'Reentry with a hateful
comment', or 'No reentry'
```

In addition to counterspeech, we also experiment with HS and the dialogue (HS+Counterspeech) to predict outcomes. For dialogue, the input for <example> is "hate comment:" + <hate comment> + "counterspeech:" + <counterspeech>.

Llama 3 Finetuned LLMs can be optimized for a specific task through fine-tuning. By feeding an LLM with training data, the model learns the linguistic patterns in counterspeech to HS with different hater reentry behaviors. We use texts and

their labels in training data to construct dialogues as input to train the Llama-3.1-8B-Instruct model. Specifically, we apply the Low-Rank Adaptation (LoRA) (Hu et al., 2021) method in finetuning (details of model parameters in Appendix B). The trained models are then utilized to make predictions on the test data.

5.2 Results

Hater Reentry Prediction This task predicts whether there is hater reentry. The majority baseline is calculated based on all samples predicted to be "reentry". Table 5 shows the performance of models. We observe the following insights:

Most models perform significantly better than the baseline, except Llama 3 zero-shot learning. This suggests the reentry behavior of haters is predictable and LLM is limited in forecasting conversation outcomes. The performance of fine-tuned Llama 3 (best F1: 0.67) is also inferior to BERT (best F1: 0.80) and BERT-MTL (best F1: 0.83).

HS/counterspeech pairs have the best predictive power across models, indicating the hater's reentry is not solely dependent on the HS or the counterspeech. Haters' reentry can be better predicted with counterspeech than HS by BERT (counterspeech (0.80) vs hate (0.74)), BERT-MTL models (counterspeech (0.81) vs hate (0.78)), and Llama 3 zero-shot (counterspeech (0.49) vs hate (0.35)). It indicates that counterspeech is important in determining whether a hater will re-engage in the follow-up conversations.

Error Cause	Non-Hateful Reentry		Hateful Reentry		No Reentry		All
	FP	FN	FP	FN	FP	FN	
Rhetorical Questions	0.52	0.47	0.38	0.63	0.58	0.41	0.49
Negation	0.22	0.26	0.29	0.17	0.20	0.26	0.23
Sarcasm or Irony	0.00	0.21	0.27	0.09	0.16	0.20	0.17
Intricate Text	0.16	0.00	0.06	0.09	0.04	0.13	0.10
General Knowledge	0.10	0.06	0.00	0.02	0.02	0.00	0.01

Table 8: Distribution of common errors in the best 3-way prediction model across different classes. The values represent the proportion of error causes. FP refers to false positives, and FN indicates false negatives.

BERT-MTL achieves the best performance (F1: 0.83). The F1 score is significantly higher than the best BERT (F1: 0.80) and Llama 3 Finetuned (F1: 0.67) based on McNemar's tests ($p < 0.001$) (McNemar, 1947), which is used to compare prediction results on the same test data between two models.

Reentry Type Prediction This task predicts whether the reentry is hateful for hater reentry cases. The baseline is calculated by assigning the majority label, i.e., non-hateful, to all samples. Table 6 shows the evaluation results. All models can beat the baseline, however, the Llama 3 zero-shot learning model is very limited in predicting the reentry type (counterspeech F1: 0.53 vs baseline: 0.57). Fine-tuned Llama 3 models do not have better results than BERT and BERT-MLT models. Similar to the hater reentry prediction, the prediction of reentry type benefits from the conversational context (BERT: hate (0.71) vs counterspeech (0.73) vs pair (0.78), BERT-MTL: hate (0.80) vs counterspeech (0.80) vs pair (82), Llama 3 Finetuned: hate (0.59) vs counterspeech (0.62) vs pair (0.62)), and the differences are significant with McNemar's tests ($p < 0.01$).

3-way Response Prediction This task predicts one of the three outcomes: hateful reentry, non-hateful reentry, or no reentry. The baseline predicts all samples with "non-hateful reentry." We report the classification results and compare them with the two-stage predictor, which combines the best models, i.e., the BERT-MTL model with pairs, to predict reentry and reentry types.

Table 7 presents the results. Llama 3 zero-shot is almost random at predicting hater reactions. The predictions based on pairs are relatively better than the counterspeech and HS alone, and can slightly beat the majority baseline. HS/counterspeech pairs can best predict haters' reactions across all trained models (BERT pair F1 (0.74), BERT-MTL pair F1 (0.77), Llama 3 Zero-shot pair F1 (0.33)), Llama 3 Finetuned pair F1 (0.46)). The two-stage prediction

shows an inferior performance (weighted F1: 0.53) to the best 3-way model (weighted F1: 0.77).

6 Error Analysis

The best 3-way prediction model still makes errors in some cases. We select 200 random samples from the prediction results by the 3-way prediction model and analyze the error cause (annotation details in Appendix D). Table 8 outlines the distribution of common mistakes observed.

Rhetorical question (0.49) causes the most common mistakes. Rhetorical questions are not meant to be answered, but to express strong emotions (Suzuki and Nadamoto, 2020), which complex the model understanding. It is also a challenge for HS detection (Schmidt and Wiegand, 2017). Rhetorical questions account for a high rate of FP and FN across all the classes. FN rate in "no reentry" (0.63) appears to be the highest, indicating that the model often misinterprets rhetorical questions in the conversation that elicit hateful reentry.

Negation (0.23) is the second most frequent error cause. It often changes the polarity of comments subtly, which is hard for the model to capture. Some HS detection models also struggle with it (Röttger et al., 2020). Negation frequently occurs when the model falsely predicts a "hateful reentry" (0.29), suggesting that the model may be oversensitive to specific keywords associated with hostility but ignore the negation elements.

Sarcasm and irony (0.17) are also challenging cases similar to other relevant tasks (Grolman et al., 2022). Users may use a humorous or satirical tone to express opinions laced with implicit hate (Frenda et al., 2022). There is a high FP rate (0.27) in "hateful reentry," similar to negation cases, suggesting the model may ignore the reversal meaning in the conversation implied by sarcasm and irony.

Our model also faces challenges with intricate texts (0.10), which use complex syntactic structures to obscure the true intent. In such cases, the

model often misclassifies the input as "non-hateful reentry" (0.16). Additionally, a few errors occur when the model lacks general knowledge (0.01), such as "incel", or "bigot," which may be easier for humans but is challenging for the model.

7 Conclusion and Discussion

We introduce a novel way to evaluate different impacts of counterspeech by analyzing hater reactions. We have curated the ReEco dataset, consisting of HS/counterspeech and the corresponding hater reactions. Our methodology involves linguistic analysis and developing models for predicting hater reactions following counterspeech, which are underexplored in previous research.

We find that certain linguistic features, such as more 2nd person pronouns, and negative words, are tied with hater reentry. Experiments show incorporating conversation contexts significantly improves the prediction of haters' reactions. The 3-way prediction method achieves better results than the best two-stage model. Our results also show that LLMs have limitations in predicting haters' reactions. Even fine-tuned LLMs are inferior to BERT and BERT-MTL models.

Our work can be applied practically. First, we provide the ReEco dataset for researchers to analyze the impact of counterspeech on haters' reactions. This dataset includes user-generated counterspeech with various conversation outcomes, which could be used to fine-tune counterspeech generation models, improving their effectiveness in managing hater behavior. Second, our linguistic findings may guide users and counterspeech generation models on how to respond to HS that are less likely to provoke further negative behavior of the hater. Third, we provide insights for social media platforms to assess and intervene in HS.

Limitations

Our study has several limitations. First, we focus on the direct hater reentry following counterspeech in triple-turn conversations, while haters' reactions in the subsequent progression of the conversation were not considered. In future work, we will investigate and categorize more nuanced hater reentry behaviors for a more comprehensive understanding of haters' reactions to counterspeech. Also, the study focuses on the immediate reactions of haters in the conversation, it would also be interesting to examine the impact of counterspeech on

haters in the long term. Second, models only use HS/counterspeech pairs for predictions. The reentry behavior is also related to users' communication habits, stances, and other conversation contextual factors such as the influence of counterspeakers (He et al., 2021), which are worthy of further investigation. Third, in our task of reentry type prediction, we only consider two types of hater reentry. A promising line of future work is to consider the analysis of reentry behaviors with a finer granularity.

Ethics Statement

We ensure that our study adheres to ethical guidelines by carefully evaluating associated risks and benefits. We collect data from Reddit [5] under Reddit's Terms of Service using PushShift API. Reddit is a public forum. When users sign up to Reddit, they consent to make their data available to the third party including the academy. Therefore, we can use Reddit data without further seeking user consent following the ethical rules (Procter et al., 2019). We have masked users' identifiable information before analysis and modeling. We will make sure the dataset is exclusively used for non-commercial research purposes[6]. Examples presented in this study may present toxic languages. We have masked such languages and made notes to readers that they may encounter toxic languages. This study aims at effective counter-speech to HS. We acknowledge the potential risks of users being re-identified with anonymized data or misuse of the data by individuals, but the benefits will outweigh such risks.

Acknowledgements

References

Dhafar Hamed Abd, Ayad R Abbas, and Ahmed T Sadiq. 2021. Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics*, 10(1):283–289.

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020.

---

[5]https://pushshift.io/api-parameters/
[6]https://www.reddit.com/wiki/api-terms/

Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Doris E Acheme, Chris Anderson, and Claude Miller. 2024. The effects of language features and accents on the arousal of psychological reactance and communication outcomes. *Communication Research*, page 00936502241229883.

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Finding authentic counterhate arguments: A case study with public figures. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13862–13876.

Alain Auger and Jean Roy. 2008. Expression of uncertainty in linguistic data. In *2008 11th International Conference on Information Fusion*, pages 1–8. IEEE.

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mfizil. 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*, pages 13–22.

Fabienne Baider. 2023. Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, 11(2):249–260.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, pages 1134–1145.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049.

Anna M Borghi et al. 2019. Linguistic relativity and abstract words. *Paradigmi*, 37(3):429–448.

Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. *Philosophy Compass*, 18(1):e12890.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754.

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Jihyang Choi and Jiyoung Lee. 2021. "enthusiasm" toward the other side matters: Emotion and willingness to express disagreement in social media political conversation. *The Social Science Journal*, pages 1–17.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation. *arXiv preprint arXiv:2307.04761*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, 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.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLIC-it 2020)*, volume 2769.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

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*, 49:803–821.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 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, pages 512–515.

Jim Dillard and Cindy Harmon-Jones. 2002. A cognitive dissonance theory perspective on persuasion. *The persuasion handbook: Developments in theory and practice*, page 99.

Rehab Duwairi, Amena Hayajneh, and Muhannad Quwaider. 2021. A deep learning framework for automatic detection of hate speech embedded in arabic tweets. *Arabian Journal for Science and Engineering*, 46:4001–4014.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 8–14. IEEE.

Seth Green, Megan Stiles, Katherine Harton, Samantha Garofalo, and Donald E Brown. 2017. Computational analysis of religious and ideological linguistic behavior. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, pages 359–364. IEEE.

Edita Grolman, Hodaya Binyamini, Asaf Shabtai, Yuval Elovici, Ikuya Morikawa, and Toshiya Shimizu. 2022. Hateversarial: Adversarial attack against hate speech detection algorithms on twitter. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 143–152.

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.

Jeffrey T Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. 2008. I'm sad you're sad: emotional contagion in cmc. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 295–298.

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*, pages 90–94.

M Honnibal and I Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. neural machine translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 688–697.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the 2018 World Wide Web Conference*, pages 1145–1154.

Nathan Lambert, Kristofer Pister, and Roberto Calandra. 2022. Investigating compounding prediction errors in learned dynamics models. *arXiv preprint arXiv:2203.09637*.

Larissa Leonhard, Christina Rueß, Magdalena Obermaier, and Carsten Reinemann. 2018. Perceiving threat and feeling responsible. how severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on facebook. *SCM Studies in Communication and Media*, 7(4):555–579.

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

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.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive emotion elicitation in chat-based dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):866–877.

George E Marcus, Michael MacKuen, and W Russell Neuman. 2011. Parsimony and complexity: Developing and testing theories of affective intelligence. *Political Psychology*, 32(2):323–336.

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 international AAAI conference on web and social media*, volume 13, pages 369–380.

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

Saif M Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 323–379. Elsevier.

Prerna Nadathur and Sven Lauer. 2020. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: a journal of general linguistics*, 5(1).

Rob Procter, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matt Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. *arXiv preprint arXiv:1908.11732*.

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.

Louis Reynolds and Henry Tuck. 2016. The counter-narrative monitoring & evaluation handbook. *Institute for Strategic Dialogue*.

Julian Risch and Ralf Krestel. 2020. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 579–589.

Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2016. Predicting news popularity by mining online discussions. In *Proceedings of the 25th international conference companion on world wide web*, pages 737–742.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. *arXiv preprint arXiv:2205.04304*.

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.

Sarah Shugars and Nicholas Beauchamp. 2019. Why keep arguing? predicting engagement in political conversations online. *Sage Open*, 9(1):2158244019828850.

Wolfgang Stroebe. 2008. Strategies of attitude and behaviour change. *Introduction to social psychology. Oxford, UK: Blackwell*.

Rinji Suzuki and Akiyo Nadamoto. 2020. Extracting rhetorical question from twitter. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, pages 290–299.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.

Luk Van Mensel, Mieke Vandenbroucke, Robert Blackwood, Ofelia García, Nelson Flores, and Massimiliano Spotti. 2016. The oxford handbook of language and society.

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*. 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.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Sebastian Wachs, Ludwig Bilz, Alexander Wettstein, Michelle F Wright, Norman Krause, Cindy Ballaschk, and Julia Kansok-Dusche. 2022. The online hate speech cycle of violence: Moderating effects of moral disengagement and empathy in the victim-to-perpetrator relationship. *Cyberpsychology, Behavior, and Social Networking*, 25(4):223–229.

Lingzhi Wang, Xingshan Zeng, Huang Hu, Kam-Fai Wong, and Daxin Jiang. 2021. Re-entry prediction for online conversations via self-supervised learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2127–2137.

Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What makes online communities 'better'? measuring values, consensus, and conflict across thousands of subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1121–1132.

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.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2023. Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech. *arXiv preprint arXiv:2312.04804*.

Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019. Joint effects of context and user history for predicting online conversation re-entries. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2809–2818.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.

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.

A   Computing Resources

The neural model takes about half an hour on average to train in a server with an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 graphic cards.

B   Hyperparameters of Models

The hyperparameters for transformer-based models are shown in Table 9. For Llama-3.1-8B-Instruct zero-shot, we set the temperature as 1.0 and use the default of other hyperparameters. The hyperparameters for all tasks with fine-tuned Llama-3.1-8B-Instruct are set to be the same. The num_train_epochs is 1, learning rate is $1e^{-5}$, train_batch_size is 1, warmup_ratio is 0.1, lora_alpha is 32, lora_dropout is 0.05, and bias is none for the training process, and the top_k is 50, top_p is 0.85, do_sample is true, and repetition_penalty is 1.0 for the inference process.

Task	Epoch	Batch size	Learning rate	Dropout
BERT-base				
Hater reentry	6	16	$1e^{-5}$	0.1
Reentry type	6	16	$1e^{-5}$	0.1
3-way	5	16	$1e^{-5}$	0.1
BERT-MTL				
Hater reentry	7	16	$1e^{-5}$	0.1
Reentry type	7	16	$1e^{-5}$	0.1
3-way	5	16	$1e^{-5}$	0.1

Table 9: Hyperparameters of transformer-based models.

C   Linguistic Factors

We employ SEANCE to conduct linguistic analysis in section 4 on corpus analysis. For most of the linguistic factors, we draw on the explanation provided by Crossley et al. (2017). To further clarify our findings, we offer a detailed definition of most linguistic factors and references we use to support our results in Table 10.

Factor Type	Definition
Textual Factors	
Uncertainty	Words denote feelings of uncertainty (Auger and Roy, 2008).
Abstract	Words indicate a tendency to use abstract vocabulary (Borghi et al., 2019).
Format	Words refer to formats, standards, or conventions of communication.
Causation	Words denote the presumption that the occurrence of one phenomenon is necessarily preceded by another (Nadathur and Lauer, 2020).
Emotional Factors	
Enlightenment	Words likely reflect a gain in enlightenment through thought, education, etc.
Polarity	Words show positive or negative sentiment, such as approval, disapproval, or neutrality (Abd et al., 2021).
Valence	Words describe intrinsic emotions, which can be either positive, negative, or neutral (Mohammad, 2021).
Longing	Words express deep yearning and strong desire.
Exclamation	Words convey strong sentiments, like surprise, and excitement.
Aggression	Words indicate anger, frustration, or hostility.
Social-related Factors	
Respect	Words demonstrate politeness, like please, thank you, and so on.
Power	Words convey authority, influence, or control in communication (Van Mensel et al., 2016).
Worship	Words express adoration or reverence, usually in a religious context (Green et al., 2017).
Forgiveness	Words refer to compassion or asking for forgiveness.

Table 10: Definitions of linguistic factors.

D   Annotation Details

Our study involves two annotation tasks: HS/ counterspeech identification in Section 3 and error analysis in Section 6. The annotation details in this study are outlined in Table 11 and Table 12. We hire two PhD students with expertise in HS/ counterspeech scope to label the data. Annotators were compensated on average with $15 per hour. We paid them regardless of whether we accepted their work. Annotators' IDs are not included in the dataset. We provide a clear taxonomy: **HS** refers to expressions in which the author deliberately targets an individual or group with the intent to vilify, humiliate, or incite hatred (Yu et al., 2022). **Other comments** refer to statements that do not exhibit such harmful intent, including neutral, supportive, or general discourse. **Counterspeech** refers to responses to hate speech that are intentionally crafted to contradict and challenge hateful remarks (Chung et al., 2023). **Other replies**, in contrast, encompass responses that do not directly counter hate speech, including neutral, supportive, or unrelated remarks.

In terms of error analysis, we refer to examples from Yu et al. (2023) to instruct annotators. Then 200 instances are randomly selected for validation. Each annotator labels the comments in the same condition but separately. After obtaining the results, the agreement rate and Kappa coefficient are calculated to prove the credibility of human annotation.

Task	HS/counterspeech Identification
Annotator Selection	Two PhD students with expertise in HS/counterspeech
Rule	**HS:** Given a comment, annotators determine whether it is HS or not, label "1" as HS, otherwise label "0".
	Counterspeech: Given a HS and its reply, annotators determine whether it is counterspeech or not, label "1" as counterspeech, otherwise label "0"
Example	**HS:** *"Then focus on those other interests? ... incel logic."*
	Counterspeech: *"You're not making any sense. I never said that, that's you saying that."*
Credibility	**HS:** Agreement rate: 93%, Kappa coefficient: 0.76;
	Counterspeech: Agreement rate: 95%, Kappa coefficient: 0.79.

Table 11: Annotation details of HS and counterspeech detection.

Task	Error Analysis
Annotator Selection	Two PhD students with expertise in HS/counterspeech
Rule	Annotators read the HS and its counterspeech, and label it as "0" for the rhetorical question, "1" for sarcasm or irony; "2" for negation, "3" for general knowledge, and "4" for intricate texts based on error examples.
Example	**Rhetorical Question:** HS: *Just because you can't be a racist s*** on Twitter doesn't mean anyone is discriminating against you.* Counterspeech: *Why would you assume anything about race?*
	Negation: HS: *You are misogynistic. If this is how you speak to a stranger on the internet, I can only imagine how nasty you are in real life.* Counterspeech: *In reality, it isn't—calling you a worthless b*** isn't misogynistic. [...] You are not "womankind."*
	Sarcasm or Irony: HS: *You are misogynistic. If this is how you speak to a stranger on the internet, I can only imagine how nasty you are in real life.* Counterspeech: *In reality, it isn't—calling you a worthless b*** isn't misogynistic. [...] You are not "womankind."*
	Intricate Text: HS: *You cannot truly see it from the other perspective either, so please take your half-ass.* Counterspeech: *How can I not see it from your point of view, and you are afraid random women will call you a rapist. That is EXACTLY THE SAME MINDSET AS FEMINISTS WHO hate all men.*
	General Knowledge: HS: *Your momma never told you not to stick your d*** in crazy?* Counterspeech: *Incel.*
Credibility	Agreement rate: 91%; Kappa coefficient: 0.75

Table 12: Annotation details of error analysis.

E   Subreddit List

We collect Reddit data from the following 42 subreddits and categorize them into five communities: Discussion, Hobby, Identity, Meme, and Media-sharing (Table 13).

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

Table 13: Subreddit list.