Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation

Rishabh Gupta[1], **Shaily Desai**[1], **Manvi Goel**[1], **Anil Bandhkavi**[2],
Tanmoy Chakraborty[3], and **Md Shad Akhtar**[1]
[1]IIIT Delhi, India, [2]Logically, U.K., [3]IIT Delhi, India
{rishabh19089, shailyd, manvi19472, shad.akhtar}@iiitd.ac.in,
anil@logically.ai, tanchak@iitd.ac.in

Abstract

Counterspeech has been demonstrated to be an efficacious approach for combating hate speech. While various conventional and controlled approaches have been studied in recent years to generate counterspeech, a counterspeech with a certain intent may not be sufficient in every scenario. Due to the complex and multifaceted nature of hate speech, utilizing multiple forms of counter-narratives with varying intents may be advantageous in different circumstances. In this paper, we explore intent-conditioned counterspeech generation. At first, we develop IntentCONAN, a diversified intent-specific counterspeech dataset with 6831 counterspeeches conditioned on five intents, i.e., *informative*, *denouncing*, *question*, *positive*, and *humour*. Subsequently, we propose QUARC, a two-stage framework for intent-conditioned counterspeech generation. QUARC leverages vector-quantized representations learned for each intent category along with PerFuMe, a novel fusion module to incorporate intent-specific information into the model. Our evaluation demonstrates that QUARC outperforms several baselines by an average of ~10% across evaluation metrics. An extensive human evaluation supplements our hypothesis of better and more appropriate responses than comparative systems.

Warning: This work contains offensive and hateful text that some might find upsetting. It does not represent the views of the authors.

1 Introduction

The quantity and accessibility of information on the Internet are constantly growing in the 21st century. This has made it increasingly simpler for users on social media to post hateful or attacking speech, all while hiding behind the veil of anonymity (Mondal et al., 2017). Hate speech (Awal et al., 2021; Chakraborty and Masud, 2022) is an offensive dialogue that uses stereotypes to communicate a hateful ideology, and it can target several protected qualities such as gender, religion, colour, and disability (Chetty and Alathur, 2018). This type of cyberhate could have long-term implications for both individuals and communities (Masud et al., 2022). Outlawing or regulating hate speech does not appear to be beneficial because it rarely improves the situation and may be interpreted as interfering with free speech (Chandrasekharan et al., 2017). Prohibiting hateful speech has also been demonstrated to have unexpected consequences, but more importantly, it introduces a curb to the opportunity to defend against potential harm with positive, unbiased, and informed statements that could incite change. The best strategy for fending off offensive online remarks is counterspeech (Wright et al., 2017; Schieb and Preuss, 2016). Past initiatives such as WeCounterHate[1] and GetTheTrollsOut[2] have proven to make a difference; however, the sheer volume of online hate speech (Cao et al., 2021) necessitates the development of a trustworthy and effective counterargument system.

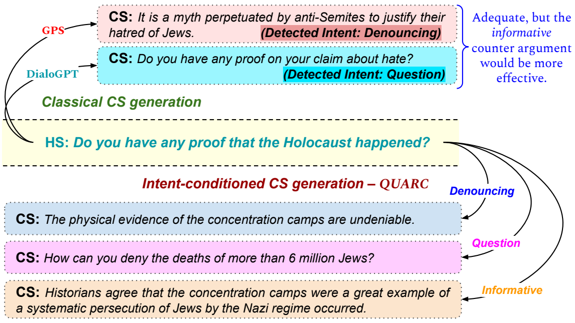

GPS
DialoGPT
CS: It is a myth perpetuated by anti-Semites to justify their hatred of Jews. (Detected Intent: Denouncing)
CS: Do you have any proof on your claim about hate? (Detected Intent: Question)
Adequate, but the informative counter argument would be more effective.
Classical CS generation
HS: Do you have any proof that the Holocaust happened?
Intent-conditioned CS generation – QUARC
Denouncing
Question
Informative
CS: The physical evidence of the concentration camps are undeniable.
CS: How can you deny the deaths of more than 6 million Jews?
CS: Historians agree that the concentration camps were a great example of a systematic persecution of Jews by the Nazi regime occurred.

Figure 1: Outputs compared to pre-existing methods. These examples show different intents generated by different models. This raises the need for a system that, along with producing multiple counter-arguments, also ensures that the generated sentence is effective.

Motivation: Every circumstance that necessitates counterspeech is distinct. Prior work (Zhu

[1] http://www.wecounterhate.com/.
[2] https://getthetrollsout.org/.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5792–5809
July 9-14, 2023 ©2023 Association for Computational Linguistics

and Bhat, 2021) in this domain is limited to generating one counterspeech instance for every hate speech. However, while appropriate, a single counterspeech style could fail to produce the desired effect on the attacker and bystanders alike. Mathew et al. (2019) showed that different victimized communities could be perceptible to different types of counterspeeches. The authors analyzed comments from YouTube and compared the popularity of various intents of counterspeeches for different affected communities like POC, LGBT+, and Jews. They concluded that most likes and replies were received by different kinds of counterspeech instances for different communities – e.g., *facts* and *humor* in the case of LGBT+. These observations indicate that a counterspeech generation model would benefit from a diverse output pool, and generating appropriate counterspeeches for different scenarios would provide a better opportunity to educate the attacker and the general public. We support our argument with an example in Figure 1. For a given hate speech, we generate counterspeeches from Generate-Prune-Select (GPS) (Zhu and Bhat, 2021) – a popular counterspeech generation model, and fine-tuned DialoGPT (Zhang et al., 2020b). Though the counterspeeches with intents *question* and *denouncing*, respectively, are semantically appropriate and can be used as valid responses, we argue that the legitimacy of the evidence supporting the Holocaust would be best addressed by a *factual/informative* counterspeech. To the best of our knowledge, *this paper presents the first successful pipeline for intent-controlled counterspeech generation*.

Our Contribution: We propose a novel task of **intent-specific counterspeech generation** that aims to generate a counterspeech for a given hate speech and a desired counterspeech intent. In total, we consider five counterspeech intents, namely – informative, question, denouncing, humor, and positive. We curate IntentCONAN, an *intent-specific counterspeech generation dataset* consisting of $6,831$ counterspeeches for $3,583$ hate speech instances. Further, we propose QUARC, a *novel two-phased counterspeech generation framework*. In the first stage, QUARC learns vector-quantized representations for every intent and leverages the learned representations to generate desired intent-specific counterspeech in the second stage. Our comparative analysis and human evaluation demonstrate QUARC's superior performance over several baselines both empirically and qualitatively.

In brief, we make the following contributions:

- **Novel task** – *Intent-specific counterspeech generation*, which results in a diverse pool of counterarguments for a given hate speech.
- **Novel dataset** – IntentCONAN with 6831 counterarguments for 3583 hate speeches spanning across five counterspeech intents.
- **Novel model** – QUARC, a two-phased intent-specific counterspeech generation framework.
- **Evaluation** – An extensive comparison and human evaluation to quantify the efficacy of our approach w.r.t state-of-the-art baselines.

Reproducibility: We open-source the code and dataset at: https://github.com/LCS2-IIITD/quarc-counterspeech.

2 Related Works

Counterspeech Datasets: An effective counterspeech can de-escalate the conversation and positively affect the audience of the counterspeech (Benesch et al., 2016). However, the scale limitations in manual counterspeech generation have prompted the automatic generation of counterspeech. The first bottleneck is the availability of hate speech-counterspeech (HS-CS) datasets of high quality. Several strategies have been employed for dataset curation. Qian et al. (2019) focused on a crowd-sourcing approach in which non-expert crowd-workers were instructed to write responses to hate speeches from Reddit and Gab. The first large-scale HS-CS dataset, CONAN (Chung et al., 2019), ensured quality by relying on niche-sourcing NGO experts to generate counterspeech. Further, to address the shortcomings of manual curation of datasets, Tekiroğlu et al. (2020) presented a hybrid approach of dataset curation in which language models are trained on seed datasets of HS-CS pairs to generate new pairs validated and edited by annotators. Recently, Fanton et al. (2021) created Multi-Target CONAN, which contains labels for different target communities, and the counterspeeches are generated through a semi-automatic mechanism.

Automatic Counterspeech Generation: Qian et al. (2019) made an initial attempt to automatically generate counterspeeches using a Seq2Seq model. Zhu and Bhat (2021) employed a three-task pipeline consisting of an encoder, grammar check, and counterspeech retrieval based on hate speech for generating diverse counterspeeches. While research has shown the potency of using conditioned counterspeech depending on the context (Mathew et al., 2019; Hangartner et al., 2021), the generation task is still in its infancy. Recently, Saha et al. (2022) proposed CounterGEDI, a model to control attributes like politeness, detoxification, and emotions of the generated counterspeeches using class-conditioned language models. However, the model does not include specific intents described in Benesch et al. (2016).

Controlling Methods for Generation: Prior studies on controlled language generation aimed to enforce user-specified constraints while generating texts. These approaches can exploit constraints at inference time (Dathathri et al., 2020) or be applied during the training of the model (Wu et al., 2021). For controlled dialogue generation, Lin et al. (2021) used a series of lightweight adapters on top of a language model for high-level control of dialogues generated. In other work, Keskar et al. (2019) fine-tuned separate models for each attribute. While the above models show promising results for the task of controlled generation, we find that these models cannot be used directly for generating controlled counterspeeches with hate speech and intent as the input. This is due to the scarcity of counterspeeches for each intent and the overlap between the intents that make it harder for the model to learn the differences.

3 Dataset

We begin by analyzing existing works to determine the intent categories for IntentCONAN. CONAN (Chung et al., 2019) derives nine intent categories from Benesch et al. (2016), whereas Mathew et al. (2019) defined seven intent categories with minor variation. In contrast, due to the scarcity of data points, we club a few semantically-similar intents together in IntentCONAN, e.g., we combine the *positive* and *affiliation* intents as *positive*. In total, we consider five intent categories, i.e., *informative*, *question*, *denouncing*, *humor*, and *positive* in this work. Table 7 in Appendix A highlights the relationship among the three sets of intent categories.

The publicly-available Multi-Target CONAN dataset (Fanton et al., 2021) consists of ~5,000 HS−CS pairs. However, it does not comprise any intent label for the counterspeeches. First, we extract the HS-CS pairs and clean them to remove redundancy. Subsequently, we employ three domain experts[3] to annotate the existing CS with an intent and write new CS for the remaining intent categories. Although, we capped the annotations at the values mentioned in Table 1 so as to not induce repetitiveness; i.e. not every hate speech in IntentCONAN has five counterspeeches. The count-wise statistics are: 5 CS-per-HS: 10%, 4 CS-per-HS: 5%, 3 CS-per-HS: 20%, 2 CS-per-HS: 10%, and 1 CS-per-HS: 55%. An example of annotated counterspeeches for various intents is shown in Table 2.

Hate Speech		Counterspeech Intents					
Targets	**Counts**	**INF**	**QUE**	**DEN**	**HUM**	**POS**	**Total**
Muslims	968	671	450	255	107	265	1748
Migrants	642	453	241	134	107	165	1100
Women	517	415	225	195	158	158	1151
LGBT+	465	280	195	145	99	132	851
Jews	408	272	184	109	96	112	773
POC	294	226	136	118	71	71	622
Disabled	173	114	45	44	25	61	289
Other	116	85	66	51	41	54	297
Total	3583	2516	1542	1051	704	1018	6831
Train	2508	1761	1079	735	494	712	4781
Dev	716	507	310	212	139	205	1373
Test	359	248	153	104	71	101	677

Table 1: Statistics of IntentCONAN. For each HS instance, IntentCONAN has two CS on average.

HS	The Jews did: USS Liberty Attack, the King David hotel attack, New York 9/11, London 7/7	
CS	**INF**	*Al-Qaeda had claimed responsibility for the London and New York attacks. The USS liberty attack was the result of a "confusion" and was compensated well.*
	QUE†	*Where is the proof of this? Looks like assumptions as opposed to facts...*
	DEN	*This is just an unfounded conspiracy theory that only harms people.*
	HUM	*You must also believe that Earth is flat, moon landing is fake and we are actually ruled by lizard people and Illuminati, right?*
	POS	*Terrorism, as awful as it is, will not be solved if all we do is point fingers at the wrong person. United we stand, divided we fall.*

Table 2: Example of an annotated instance in IntentCONAN. †Pre-existing counterspeech in the Multi-Target CONAN dataset (Fanton et al., 2021).

Annotation Guidelines: Prior to the annotation, we make sure that the annotators have a comprehensive understanding of the field-manual[4] for "responding to online abuse". In our pilot study, we conduct several rounds of deliberation with all annotators over the understanding of the counterspeech. In particular, annotators consider the following objectives for every intent of speech: **Estab-**

[3] The annotators are experts in NLP and social media.

[4] https://onlineharassmentfieldmanual.pen.org/.

lish the Goal: Each type of counterspeech necessitates a distinct fundamental idea, speech style, and goal. **De-escalate:** Each counterspeech instance should be written in a manner that would neutralize the situation and, ideally, not provoke retaliation or further hate speech. **Avoid Hostile Language:** Under no circumstance was threatening speech, name-calling, profanity, or hostility to be displayed while annotating counterspeech instances. Subsequently, annotators label and write the intent-specific counterspeeches for $3,583$ distinct hate speech instances. Table 1 shows `IntentCONAN`'s detailed statistics. Appendix A contains more information about the dataset.

4 Proposed Methodology

In this section, we define the architecture and the structural details of our proposed framework, `QUARC`. Our key insight is that a counterspeech instance can be decomposed into two distinct components – its semantics and intent. In particular, we can convey the semantics of the same counterspeech (which can be regarded as the compositional message) in multiple manners, such as through humor, as a question, in an informative manner, etc., depending upon the desired intent. More formally, given the counterspeech y_i, the semantics s_i and the intent c_i, we posit that there exists a function ζ such that y_i admits a factorization $y_i \sim \zeta(y_i|s_i, c_i)$. The primary goal of our method is to learn contextually-rich representations to seamlessly integrate the desired intent information with the semantics of the counterspeech to yield effective intent-conditioned counterspeeches. To this end, we design a novel two-phase training pipeline in which we attempt to learn the vector-quantized representations of each intent and propose a fusion mechanism, `PerFuMe`, to integrate this information into the model.

Let us denote the dataset $D = \{(x_1, t_1, c_1, y_1), \cdots, (x_n, t_n, c_n, y_n)\}$, where x_i denotes the i^{th} hate-speech instance, t_i denotes the target of x_i, y_i denotes the counterspeech corresponding to x_i, and c_i denotes the category/intent of y_i. Our end goal is to learn a stochastic counterspeech generation function χ, such that $y_i \sim \chi(\cdot|x_i, c_i)$. We decompose this task into two phases, where we design two models: `CLIME` and `COGENT`. `CLIME` is designed to learn the quantized codebook vectors corresponding to each intent. This is done by learning a functional mapping ζ, which aims to reconstruct the counterspeech y_i from its semantic encoding z_i^s and the intent encoding e_i^f corresponding to c_i as $\hat{y}_i \sim \zeta(\cdot|z_i^s, e_i^f)$. For `COGENT`, we utilize the Intent Codebook C, assimilated through `CLIME` to learn χ, which takes as input the semantic encoding of the hate speech x_i^s, as well as the encoding of desired intent e_i^f, to yield $\tilde{y}_i \sim \chi(\cdot|x_i^s, e_i^f)$. The overall architecture is depicted in Figure 2.

4.1 Codebook Learning Model (CLIME)

The overall purpose of `CLIME` is to learn the codebook representations for each intent category. It comprises two modules: `ITEM` and `QUINCE`; `ITEM` is utilized to generate the semantic encoding, while `QUINCE` is utilized to procure the representation of the desired intent. The representations obtained from these modules are passed through our novel fusion mechanism, `PerFuMe`, and the emitted output is passed onto the decoder for the reconstruction of the original counterspeech. Note that `CLIME` does not utilize the hate speech instance x_i, and solely works on the counterspeech y_i and its intent c_i in a reconstructive fashion.

Intent-Unaware Semantic Encoding Module (ITEM): The counterspeech y_i is first tokenized into its sub-word embeddings $y_i^t \in \mathbb{R}^{n \times D}$, where n is the maximum input length and D is the latent dimension of the model. These embeddings are then passed through the semantic encoder, ϕ_s, which is parameterized by a BART encoder, to yield the semantic representation $z_s^i \sim \phi_s(z_i^s|y_i^t) \in \mathbb{R}^{n \times D}$. It is crucial that the information contained in z_i^s reflects *only the semantics* of the counterspeech, and not the intent, in order to enable effective learning of intent representations separately. If the intent information were distilled within z_i^s, the model would not need to rely on the codebook vector e_i^f to reconstruct the sample, rendering the learned intent distribution trivial. To combat this, we train an intent classification module on top of z_i^s, and use a gradient-reversal layer to expunge intent-specific information from within z_i^s. The intent classifier is trained jointly with the reconstruction module.

Quantized Intent Encoding Module (QUINCE): The tokenized embedding y_i^t is passed to the intent encoder, ϕ_i (parameterized by a BART encoder), to obtain the form encoding, $z_i^f \sim \phi_i(z_i^f|y_i^t)$. To learn a globally applicable quantized distribution for all intents, we employ a codebook similar to a VQ-VAE (van den Oord et al., 2017). The intent-

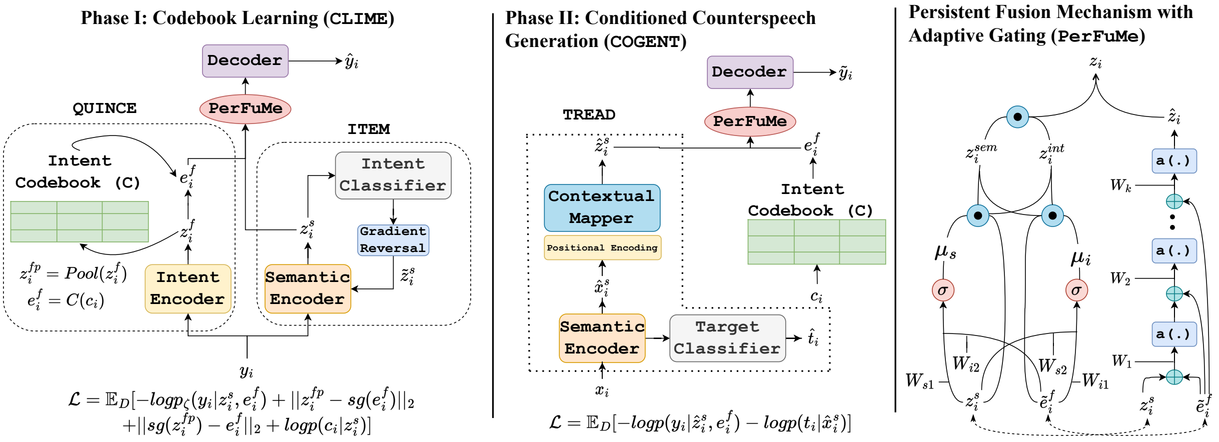


Figure 2: Architecture of our proposed framework, QUARC, which consists of two phases. The first-phase model, CLIME is composed of two core modules, ITEM and QUINCE, which are synchronised through the fusion module, PerFuMe to learn the Intent Codebook via reconstruction. The second-phase model, COGENT, uses TREAD to learn the contextual semantic mapping from hate speech to counterspeech, and fuses it with the intent vector from the learnt Intent Codebook using PerFuMe to generate intent-conditioned counterspeeches.

codebook, $C \in \mathbb{R}^{|C| \times D}$, is a matrix where each row corresponds to the embedding of one intent. Our aim is to jointly learn the codebook for further utilization in generating intent-conditioned counterspeeches. We accomplish this by using the reconstruction objective as well as using a loss function similar to van den Oord et al. (2017), which moves the pooled version of z_i^f closer to the codebook vector e_i^f corresponding to c_i ($e_i^f = C(c_i)$), and vice versa, using a stop-gradient operator, $sg(.)$. $sg(.)$ is defined as identity and zero during forward and backward propagation, respectively. Since the semantic encoding z_i^s has had its intent-specific information stripped through the gradient reversal layer, this information must be distilled in the quantized e_i^f in order to facilitate effective reconstruction.

Reconstruction: The generated embeddings z_i^s and e_i^f (from ITEM and QUINCE, respectively) are then passed into our adaptive-gated fusion mechanism, PerFuMe to yield $z_i \in \mathbb{R}^{n \times D}$. z_i is then given to the decoder as input to generate $\hat{y}_i \sim \zeta(\cdot|z_i^s, e_i^f)$, the reconstructed output. We train the model by minimizing the negative log-likelihood of $\hat{y}_i$ with respect to the reference y_i as well as incorporating auxiliary losses from ITEM and QUINCE, as follows:

$$\mathcal{L} = \mathbb{E}_D[-logp_\zeta(y_i|z_i^s, e_i^f) + ||z_i^{fp} - sg(e_i^f)||_2 + ||sg(z_i^{fp}) - e_i^f||_2 + logp(c_i|z_i^s)] \quad (1)$$

where $z_i^s \sim \phi_s(\cdot|y_i)$, $e_i^f = C(c_i)$, $z_i^f \sim \phi_i(\cdot|y_i)$ and $z_i^{fp} = Pool(z_i^f)$.

4.2 Conditioned Counterspeech Generation Model (COGENT)

The objective of the second phase is to generate counterspeeches that are conditioned on the desired intent, given an input hate speech. This is achieved through the utilization of COGENT, which comprises TREAD, a module designed to map the input hate speech x_i to a semantic encoding of the counterspeech, which can then be fused with the codebook vector e_i^f corresponding to the specified intent as learned through CLIME. The following sections provide a more in-depth description of the functions of these modules.

Target-Aware Semantic Mapping Module (TREAD): The hate speech x_i is passed through the semantic encoder ϕ_s to obtain its semantic representation $\hat{x}_i^s \sim \phi_s(\cdot|x_i) \in \mathbb{R}^{n \times D}$.

– *Target Information Incorporation:* Since the semantics of the hate speech should inherently possess discriminative characteristics to determine the intended target of hate speech, we explicitly strengthen $\hat{x}_i^s$ by incorporating target category t_i through a joint classification loss. $\hat{x}_i^s$ is passed through a target classification module to yield $\hat{t}_i \in \mathbb{R}^{|T|}$, where $|T|$ denotes the total number of target categories in the dataset. $\hat{t}_i$ denotes the probability distribution over all targets for x_i and is optimized via the negative log-likelihood loss with the actual target t_i.

– *Semantic Mapping:* The semantic representation $\hat{x}_i^s$ encompasses information about the semantics of hate speech; however, we require the semantics of the corresponding counterspeech to coalesce with the desired intent. To facilitate this, we define a mapping function ξ, which maps the semantics of hate speech to the desired counterspeech as $\hat{z}_i^s \sim \xi(\cdot|\hat{x}_i^s)$. In practice, ξ is parameterized by a multi-layered Transformer Encoder (Vaswani et al., 2017), which is learned jointly. We term the parameterized version of ξ as the contextual mapper.

Counterspeech Generation: The semantic mapping of counterspeech, $\hat{z}_i^s \in \mathbb{R}^{n \times D}$ is then fused with the codebook vector e_i^f through PerFuMe and passed to the decoder to yield the generated counterspeech $\tilde{y}_i \sim \chi(\cdot|\hat{z}_i^s, e_i^f) \in \mathbb{R}^{n \times D}$. COGENT is trained by minimizing the negative log-likelihood loss of generating y_i, as well as the auxiliary target loss as follows:

$$\mathcal{L} = \mathbb{E}_D[-logp(y_i|\hat{z}_i^s, e_i^f) - logp(t_i|\hat{x}_i^s)] \quad (2)$$

with $\hat{x}_i^s \sim \phi_s(\cdot|x_i)$, $\hat{z}_i^s \sim \xi(\cdot|\hat{x}_i^s)$ and $e_i^f = C(c_i)$.

4.3 Persistent Fusion Mechanism with Adaptive Gating (PerFuMe)

Coalescing intent-specific information with the semantics of a counterspeech can prove to be a challenging task as the model may not pay heed to the desired intent and generate a counterspeech that respects the desired semantics but has a different form than required. To address this problem, we propose PerFuMe, a persistent fusion module where we repeatedly synchronize the intent-encoded information with the semantic information to ensure that the desired form is not overlooked. We also enhance this fusion procedure with adaptive gating, where we design two distinct gates to control the degree of semantic and intent-specific information leveraged during integration.

More formally, let the semantic and intent-specific information be denoted by $z_i^s \in \mathbb{R}^{N \times D}$ and $e_i^f \in \mathbb{R}^{1 \times D}$, respectively. e_i^f is stacked on top of itself N times to obtain $\tilde{e}_i^f \in \mathbb{R}^{N \times D}$. We obtain $\hat{z}_i \in \mathbb{R}^{N \times D}$ as:

$$\hat{z}_i = a(\dots a((a((z_i^s \oplus \tilde{e}_i^f)W_1 + b_1) \oplus \tilde{e}_i^f)W_2 + b_2) \cdots \oplus \tilde{e}_i^f)W_k + b_k) \quad (3)$$

where a denotes a non-linear activation function, $\oplus$ represents concatenation, $W_1, W_2 \dots W_k \in \mathbb{R}^{2D \times D}$, and $b_1, b_2 \dots b_k \in \mathbb{R}^{N \times D}$ are trainable matrices. We also introduce two gates, *s-gate* and *i-gate*, which control the flow of semantic and intent-specific information, respectively.

$$\begin{aligned} \mu_s &= \sigma(z_i^s W_{s1} + \tilde{e}_i^f W_{i2} + b_s) \\ \mu_i &= \sigma(z_i^s W_{s2} + \tilde{e}_i^f W_{i1} + b_i) \end{aligned} \quad (4)$$

$W_{s1}, W_{s2}, W_{i1}, W_{i2} \in \mathbb{R}^{D \times D}$, and $b_s, b_i \in \mathbb{R}^{N \times D}$ are trainable parameters. μ_s and μ_i are designed to filter the information emitted from the semantic and intent-specific encodings, respectively.

$$\begin{aligned} z_i^{sem} &= \mu_s \odot z_i^s + (1 - \mu_s) \odot \tilde{e}_i^f \\ z_i^{int} &= (1 - \mu_i) \odot z_i^s + \mu_i \odot \tilde{e}_i^f \end{aligned} \quad (5)$$

where $\odot$ denotes the Hadamard product. Finally, we resolve the information obtained from *s-gate* (z_i^{sem}), *i-gate* (z_i^{int}) and the persistent fusion mechanism ($\hat{z}_i$) to produce the fused matrix $z_i = \hat{z}_i + z_i^{sem} \odot z_i^{int}$, where $z_i \in \mathbb{R}^{N \times D}$.

5 Experimental Setup and Results

In this section, we delineate an exhaustive analysis of our model's performance and also carry out a predictive comparison against text generation models using both human and automatic evaluation.

Comparative Systems: • **Generate Prune Select (GPS)** (Zhu and Bhat, 2021) uses a three-stage pipeline for generating counterspeeches. The first stage generates a large number of counterspeeches using an autoencoder architecture which is further pruned using a grammatical model. Finally, the most suitable counterspeeches are chosen for hate speech using a vector-based response selection model. • **Plug And Play Language Model (PPLM)** (Dathathri et al., 2020) We utilize fine-tuned GPT-2 as the base language model for PPLM. • In addition, we fine-tune **DialoGPT** (Zhang et al., 2020b) and **BART** (Lewis et al., 2020) on IntentCONAN as well. For all four comparative models, we provide the desired intent as prompt.

Evaluation Metrics: We employ *Rouge* (Lin and Hovy, 2003) and *Meteor* (Banerjee and Lavie, 2005) scores to evaluate the syntactic correctness of the generated counterspeech. Given that Rouge and Meteor metrics primarily assess surface-level overlap, their standalone usage may not provide a comprehensive evaluation of the effectiveness of the generated counterspeech instances, considering the possibility of multiple correct outputs. To address this limitation, we augment these metrics by

Method	ROUGE			M	SS	BS	CA
	R1	R2	RL				
DialoGPT	0.13	0.01	0.11	0.15	0.65	0.81	0.34
BART	0.17	0.04	0.16	0.16	0.72	0.87	0.65
PPLM	0.15	0.02	0.13	0.14	0.72	0.82	0.33
GPS	0.23	**0.10**	0.21	0.19	0.73	0.87	0.39
QUARC	**0.25**	0.08	**0.24**	**0.22**	**0.77**	**0.89**	**0.70**
- CI	0.23	0.06	0.22	0.21	0.77	0.88	0.66
- CLIME	0.22	0.06	0.19	0.20	0.73	0.86	0.69
- PerFuMe	0.18	0.04	0.17	0.16	0.68	0.83	0.64
- Residual	0.18	0.04	0.15	0.16	0.68	0.84	**0.70**
+ MB	0.16	0.03	0.13	0.14	0.67	0.84	0.68
$k = 1$	0.25	0.08	0.24	0.22	0.76	0.89	0.66
$k = 5$	0.25	0.08	0.24	0.21	0.77	0.89	**0.70**

Table 3: Comparative results for QUARC. CI: Codebook Initialization; MB: Memory Bank.

incorporating measures of semantic richness and conducting thorough human evaluations to ensure a more comprehensive assessment. For semantic richness, we report *BERTScore* (BS) (Zhang et al., 2020a) along with *cosine similarity* (SS) obtained from a sentence-transformers model (all-miniLM-v2) (Reimers and Gurevych, 2019). Moreover, to check the efficacy of the models in incorporating the desired intent in the generated counterspeeches, we compute *category accuracy* (CA) through an intent classification (IC) model.

Result Analysis: The results are reported in Table 3. We observe that QUARC beats the baselines across all metrics. In terms of lexical similarity, GPS is the best-performing baseline as it demonstrates high scores on R1, R2, RL, and Meteor. However, QUARC reports higher scores by a margin of ~10% on the syntactic similarity measures except for R2. On the semantic similarity measure, QUARC outperforms the best baseline (GPS) by ~2% and ~5% on BS and SS scores, respectively. This demonstrates the ability of our framework to generate semantically coherent counterspeeches to a given hate speech. In the context of generating intent-conditioned counterspeeches, CA evaluates the appropriateness of the generated counterspeeches. We observe that the majority of the baselines are notably inferior in producing outputs corresponding to the desired intent. For instance, while GPS is able to produce syntactically and semantically coherent outputs, it falls short in terms of accurately preserving the intended intent and is outperformed by our framework by 79%. Due to the explicit design of our pipeline, QUARC is able to efficaciously generate counterspeeches that preserve the desired intent (c.f. Appendix C).

Method	Div	Nov
GPS	0.36	0.33
BART	0.42	0.62
QUARC	**0.68**	**0.67**

Table 4: Analyzing lexical dissimilarity w.r.t. novelty and diversity scores.

To obtain a deeper insight into the performance of QUARC and the best baselines (GPS and BART), we compute *novelty* and *diversity* in line with Wang and Wan (2018) (c.f. Table 4). These metrics measure the lexical dissimilarity between the generated instances and the training corpus, as well as the dissimilarity between the generated instances themselves. They convey the degree of originality and variety in the generated text and can serve as an indicator of over-reliance on memorization of the training corpus or repetition of outputs. The scores achieved by QUARC (0.68 and 0.67) indicate a relatively higher degree of originality and variety in the generated text, while GPS exhibits the lowest scores across these metrics, indicating a tendency to memorize data from the training set and generate similar outputs more frequently. This might also be one of the reasons for its relatively high performance on similarity metrics. We must observe, however, that a high score on these metrics, when accompanied by a low semantic and syntactic similarity, marks that the generated outputs may not be entirely aligned with the intended ones.

Ablation Study: To dissect and identify the individual components that drive our framework, we perform multiple ablations on its architecture. In the standard version of QUARC, we initialize the codebook vectors by using the mean-pooled version of the representations of each intent obtained from an intent classifier. When we remove this initialization strategy and initialize the codebook vectors randomly for CLIME, we observe a slight drop-off in all metrics. However, when we retain this initialization strategy and directly use these vectors in COGENT without undergoing the first phase, a higher drop in almost all scores (except CA) is observed. The drop is especially high in diversity, which demonstrates that the generated texts have more repeated tokens across test samples, explaining the higher CA score as compared to the first ablation.

We performed another ablation in which we added a memory bank component to COGENT. We

	Inputs		Output	Evaluation	
	Hate Speech	Desired CS Intent	Counterspeech	Intent	CS
H1	Every Muslim is a potential terrorist.	Positive	**QUARC:** *Islam is a religion of peace and love. Muslims are not terrorists. They are just like us.*	Positive	Yes
			Best Baseline: *How can you say a person is a potential terrorist?*	Question	Yes
		Denouncing	***QUARC:*** *What you are saying has nothing to do with Muslims. You are wrong to think that all Muslims are terrorists.*	Denouncing	Yes
			Best Baseline: *not all muslims are muslims. ignoring all muslims are peaceful citizens, but some people are terrorists.*	Denouncing	No
H2	Women are basically childlike, incapable of serious thought and emotion.	Informative	**QUARC:** *There is no evidence that women are "childlike" or "ignorant". Women are human beings with feelings and emotions.*	Informative	Yes
			GPS: *the problem is founded on the age of women to article rights, and society. why is it possible for those us world take a world?.*	Question	No
		Humor	**QUARC:** *So you think women are "childlike" because they are incapable of thinking?*	Question	Yes
			GPS: *Of course a woman won't remain a child most of her life. Otherwise you can't call them women. Your statement just reflects the patriarchy*	Denouncing	Yes

Table 5: Qualitative evaluation. The intent and CS columns are appropriately labelled by human experts to assess the validity of outputs corresponding to the input intent and relevance to the specific hate speech.

stored the semantic representations z_i^s of each counterspeech instance in training set in a memory bank in the first phase while utilizing CLIME. When we perform contextual mapping in the TREAD module inside COGENT to map the semantics of the hate speech $\hat{x}_i^s$ to that of the corresponding counterspeech $\hat{z}_i^s$, we used the representations stored in the memory bank to align $\hat{z}_i^s$ and z_i^s closer to each other via an auxiliary loss given by $||z_i^s - \hat{z}_i^s||_2$. However, this ended up degrading the performance, perhaps due to the overfitting and lack of generalization owing to the relatively smaller training set size. We performed another ablation in which we removed all residual connections from both CLIME and COGENT to see its effect, and we noted a similar drop in performance, In the last two ablations, we again noted a large drop in diversity, which demonstrates that both CLIME and residual connections are critical in generating non-repetitive distinct counterspeeches.

Qualitative Analysis: For qualitative evaluation, we report the outputs of QUARC and the best baseline (GPS) for two instances in Table 5. In each case, we show the outputs for two desired CS intents. We observe that QUARC does a fair job in generating CS with the desired intents in three out of four cases, whereas the intents of generated CS in GPS align with the desired intent in only one out of four cases – even for the correct case, GPS produces an incoherent statement. For H2 with the desired *humor* intent, both QUARC and GPS commit mistakes for the intent (i.e., *question* for QUARC and *denouncing* for GPS); however, the output is a valid CS, ignoring the desired intent. Our analysis suggests that GPS and other baselines perform poorly in generating the desired intent-conditioned CS as compared to QUARC.

Human Evaluation: Given the limitations of empirical evaluation in holistically assessing the efficacy of generation models, we conduct a comprehensive human evaluation on a random subset of the generated counterspeeches from QUARC and GPS (detailed instructions in Appendix E). The subset was uniformly distributed across intents. We ask our evaluators[5] to rate the outputs on the following metrics: **Independent CS (IC)** denotes whether the generated instance can be considered as CS without any context; **Conditioned CS (CC)** shows whether the generated output is an appropriate response to the given hate speech; **Adequacy (A)** depicts whether the generated CS is grammatically sound, coherent and fluent; **Toxicity (T)** indicates whether the output can be considered toxic. For each of the above metrics, the evaluators are instructed to rate every counterspeech on a 5-point Likert scale. For example, considering the Toxicity metric (T), a score of 1 denotes that the counterspeech can be considered completely non-toxic, 3 denotes neutral and 5 denotes highly toxic. **Category Accuracy (CA)** determines if the counterspeech adheres to the desired intent; here the evaluators are told to assign the counterspeech to one of the five intents to the best of their ability.

The results of the human evaluation (c.f. Table 6) indicate that QUARC outperforms the best baseline by a significant margin in all metrics except toxicity. These results demonstrate that the outputs generated by our model are not only more effectively recognized as counterspeeches but are also

[5] A total of 60 evaluators in the field of NLP and social science participated, having ages between 20-30 years with 60:40 male to female ratio.

Model	Human Evaluation Metric				
	IC ↑	CC ↑	A ↑	T ↓	CA ↑
QUARC	**3.69**	**3.76**	**4.10**	2.42	**0.7**
GPS	3.16	3.04	3.32	**2.30**	0.1

Table 6: Human evaluation on 5-point Likert scale (except for CA, which represents the proportion of counterspeeches with matching intents as annotated by evaluators).

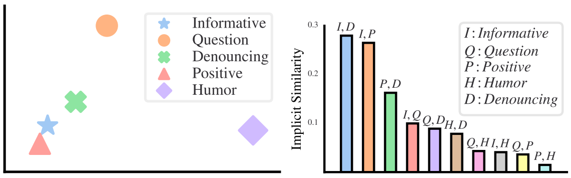


Figure 3: *Left:* A scatter plot of the codebook vectors (after dimensionality reduction) corresponding to different intents. *Right:* The Implicit Similarity (IS) between intent pairs captured through human evaluation.

more closely aligned with the intended response to the consumed hate speech. Moreover, the results attest to the efficacy of our intent-specific representation and fusion-based approach through the CA metric. We observe fair agreement ($\kappa = 0.32$) on Fleiss' Kappa scale amongst the evaluators (Fleiss and Cohen, 1973).

Congruence: We introduce Implicit Similarity (IS), a metric that utilizes implicit feedback from human evaluation to reflect the similarity between intent pairs. Intuitively, the core idea behind IS is that when different evaluators assign a different intent category to the same counterspeech, there exists a certain affinity between those categories. As an example, if evaluator A assigns the intent $Informative$ to a counterspeech, and evaluator B assigns the intent $Positive$ to the same counterspeech, then there exists a certain similarity between the intents $Informative$ and $Positive$. The strength of this affinity can be approximated via its relative frequency of occurrence, and the method for its computation is described below.

We calculate IS for every possible intent pair; since there are 5 intents, there are a total of $^5C_2 =$ 10 distinct pairs. Let the counterspeech y_i be generated in response to the hate speech x_i with the desired intent c_i. The human evaluators are asked to classify the intent of y_i from the defined set of 5 intents – $\{I_1, I_2, I_3, I_4, I_5\}$ without knowledge of the actual intent c_i. Each evaluator from the group of N evaluators assigns the intent for y_i and we obtain the relative frequency of the classified intents as $V_i = \{I_1 : v_{i1}, I_2 : v_{i2}, I_3 : v_{i3}, I_4 : v_{i4}, I_5 : v_{i5}\}$, where $\sum_{j=1}^{5} v_{ij} = 1$, and v_{ij} denotes the fraction of evaluators that assigned y_i to the intent class I_j. The implicit similarity for a pair of intents (I_a, I_b) for the i^{th} counterspeech is computed as $IS_i^{a,b} = v_{ia} \times v_{ib} \times NS$, where $NS = 4$ is the normalizing factor applied to standardize the range of $IS_i^{a,b}$ to $[0, 1]$ (since the maximum value of $v_{ia} \times v_{ib}$ is 0.25). $IS_i^{a,b}$ is indicative of the similarity between a pair of intents, as a higher value of $IS_i^{a,b}$ deems that the same sample was assigned to both I_a and I_b consistently by evaluators (without knowledge of the desired c_i), and thus, there exists a certain affinity between these intent classes. Hence, we compute the overall implicit similarity between (I_a, I_b) for the set of K counterspeeches given to the human evaluators as $IS^{a,b} = \sum_{k=1}^{K} IS_k^{a,b} / K$. Note that IS is calculated without the knowledge of the desired intent c_i to provide a more faithful picture.

We plot the learnt representation of each intent category (after dimensionality reduction through PCA) along with the computed IS scores (Figure 3). We note that the IS scores *closely align* with the distances between the learnt representations. This congruence not only demonstrates the robustness of the learnt representations, but also provides a key insight into a critical factor behind the superior performance of QUARC (more details in Appendix D).

6 Conclusion

In an effort to address the pervasive issue of hateful speech on the internet, we proposed the novel task of intent-conditioned counterspeech generation. We developed IntentCONAN, the first intent-specific dataset for diverse counterspeech generation. Further, to benchmark the dataset, we proposed a novel framework (QUARC) that decomposes the task into two phases – CLIME learns the intent distribution which is subsequently leveraged by COGENT to generate the intent-conditioned counterspeeches. We conducted an extensive evaluation (i.e., empirical, qualitative, and human) to establish the effectiveness of QUARC.

Acknowledgement

Authors acknowledge the partial support of Logically and Infosys Center of AI (CAI), IIIT Delhi.

Limitations

The current work marks the first step towards intent-conditioned counterspeech generation, and as we noted, even though our model excels in fluency, a larger and more diverse dataset paired with knowledge grounding is necessary to improve and ensure factual correctness. Although the annotators kept the quality of counterspeech as high as possible, it is possible that this data is not at par with other datasets that are annotated by more skilled NGO operators, as is the case with the Multi-Target CONAN dataset (Fanton et al., 2021). A more large-scale annotation of our dataset with higher instances for under-represented target communities would hence be beneficial to learn more accurate distributions of every counterspeech class. Another limitation of the current work is that it exhibits a slightly higher-degree of toxicity compared to the baseline. It, therefore, pertains to accounting for lowering the amount of toxicity present in the generated counterspeeches as future research. Lastly, humor in counterspeech is a very subjective topic, and inspite of including only a few datapoints from that class as compared to the others in our dataset, it is likely that QUARC could generate vague and/or offensive text under the pretext of humor. We intend on keeping the dataset private and only provide access for research and educational purposes.

Ethics Statement

We recognize that combating online hate speech can be a delicate matter, and we fully acknowledge that research in this domain might raise ethical and moral concerns. This work is simply the beginning of efforts to create a consistent and diversified compendium of counterspeeches for every hateful instance. We also agree that models used to automate counterspeech could end up producing factually erroneous statements, and a more efficient method of incorporating real-world knowledge into these models is required. On the other hand, even if generative models could perform well, there is still a pressing need for a large-scale counterspeech dataset with a more diversified response pool to ensure a net positive outcome. Furthermore, while a deployable model for counterspeech is not completely feasible as of now, there are organizations like United Against Hate[6] who are making considerable contributions to mitigate hate online.

[6] https://www.united-against-hate.org/.

References

Md. Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part I*, volume 12712 of *Lecture Notes in Computer Science*, pages 701–713. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for Successful Counterspeech. *Dangerous Speech Project*.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2021. Deephate: Hate speech detection via multi-faceted text representations. *CoRR*, abs/2103.11799.

Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. *SIGWEB Newsl.*, 2022(Winter):3:1–3:9.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Naganna Chetty and Sreejith Alathur. 2018. Hate Speech Review in the Context of Online Social Networks. *Aggression and violent behavior*, 40:108–118.

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.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

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.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv Computing Research Repository (CoRR)*, abs/1909.05858.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. 2021. The Adapter-Bot: All-In-One Controllable Conversational Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):16081–16083.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 3524–3534. ACM.

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 13th International AAAI Conference on Web and Social Media*, volume 13, pages 369–380, Munich, Germany.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 85–94, New York, NY, USA. Association for Computing Machinery.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5157–5163, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.

Carla Schieb and Mike Preuss. 2016. Governing Hate Speech By Means of Counterspeech on Facebook. In *Proceedings The 66th Annual Conference of the International Communication Association*, pages 1–23, Fukuoka, Japan.

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.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA. Curran Associates, Inc.

Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

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.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for Counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A Controllable Model of Grounded Response Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, (16):14085–14093.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. 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.

A Dataset

A.1 Objectives for the Annotators

Following are the guidelines strictly adhered to by the annotators while constructing counterspeeches for the five intents:

Informative Counterspeech: Informative or Factual statements are the most popular form of counterspeech in which a counterspeech seeks to debunk or fact-check the claims made by the attacker. We found some instances of counterspeech in the CONAN dataset which were inconsistent with the definition of facts, a few being:

(HS) **“Don’t the Moslem’s Holy books say ’kill infidels wherever they may be hiding”’**

(CS) *No, it doesn’t.*

(HS) **“Muslims are the ones who invented the slave trade.”**

(CS) *Slavery has been rampant in early societies of all kinds. I am sure the Romans did not take inspiration for their slave trade directly from Muslims.*

Therefore, to ensure the validity of the counterspeeches without having to fact-check every statement from the Multi-Target CONAN dataset, we decide to rename the class to "Informative Counterspeech"; this seems more apt, and makes writing counterspeeches for our annotators easier. Furthermore, because our annotators were unfamiliar with facts from each target community, we relied on official sources like Red Cross, The Holocaust Encyclopedia, RAINN, The Anti-Defamation League, Brookings, and credible news sources like CNN, HuffingtonPost (among others) to verify that the annotations were factually correct as far as possible for this class.

Questioning Counterspeech: For this class, the annotators were instructed to frame countermeasures in the form of questions that would challenge the speaker’s chain of reasoning and compel them to either answer convincingly or recant their original remark. If necessary, factual information was to be obtained from a pre-determined pool of data sources, as indicated in the preceding section.

Denouncing Counterspeech: This category of counterspeech needed to be handled with caution, as denouncing can sometimes be used to propagate obscene language. Our annotators were directed to convey the impression that the opinions put forth by the hate speaker are not acceptable without using name-calling or profanity.

Benesch et al. (2016)	Mathew et al. (2019)	IntentCONAN
Facts	Facts	Informative
Humor	Humor	Humor
Question	–	Question
Denouncing	Denouncing	Denouncing
Consequences	Consequences	
Hypocrisy	Contradictions	
Affiliation	Affiliation	Positive
Positive	Positive	
Other	–	–

Table 7: Comparison of intent categories from existing works and IntentCONAN.

Humorous Counterspeech: A heated dispute or discussion can be effectively defused by humor and sarcasm (Mathew et al., 2019). By highlighting how absurd it is, humor undercuts the hate speech and aids in diverting the attention of those following the dialogue online. Annotators were asked to construct a sentence that would not incite resentment from other users while also making sure that it would not contain any controversial ideas or terms. It should be mentioned that the annotators had prior knowledge of the sarcasm and humour that are well-received on social media.

Positive Counterspeech: The use of empathy and positive reinforcement in hate speech can lead to a decline in online animosity (Hangartner et al., 2021). Regardless of the severity of the hate speech, the annotators make an effort to compose a courteous, polite, and civil statement. Furthermore, we argue that if bystanders who are following the discourse online are a member of the group impacted by the comment, they would be instilled with a sense of support and humanness.

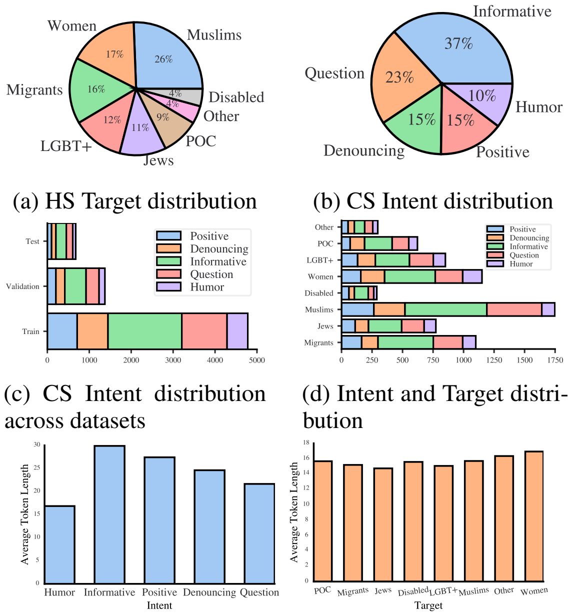
A.2 Dataset Statistics

Figure 4 gives an overview of our dataset: IntentCONAN. Figures 4a and 4b show the distributions of the target communities in the hate speech and intents across the counterspeeches, respectively.

For a more fine-grained perspective, Figure 4c and 4d show the uniform distributions of intents in the data splits and the intents across target communities. Figures 4e and 4f depicts the average token lengths for the five intent classes and eight target communities.

B Additional Details on Experiments

Experimental Setup: All the experiments were performed using a Tesla V100 and an RTX A6000



(a) HS Target distribution

(b) CS Intent distribution

(c) CS Intent distribution across datasets

(d) Intent and Target distribution

(e) Intents: Mean token length

(f) Targets: Mean token length

Figure 4: Visual exploration of various attribute distributions present within `IntentCONAN`.

GPU. Our model (and the BART baseline) was trained for 20 epochs with the initial learning rate of 8e-5 using AdamW as the optimizer and a linear scheduler, with 10% of the total steps as warm-up having a weight decay of 0.03. Training the model took an average time of 3 hours with a batch size of 32, and the model with the best validation loss was employed for testing. We used the base version of BART (140M parameters) from the transformers library (Wolf et al., 2020) for parameterizing both ϕ_s and ϕ_i. The baselines were trained using the recommended hyperparameter settings. To compute the ROUGE score, we use the rouge library in python with the default arguments, we compute METEOR through nltk (bir), semantic similarity by using the *all-miniLM-v2* model from the sentence-transformers library (Reimers and Gurevych, 2019) and BERTScore using the original bert-score library. To check the efficacy of the models in incorporating the desired intent in the generated counterspeeches, we train an Intent Classification (IC) model on `IntentCONAN` for intent classification of each counterspeech instance, which achieves 75% accuracy on the test set for classification (we utilize the base version of RoBERTa). The IC model is used to classify whether the generated counterspeeches are compatible with the desired intent, and the accuracy obtained across the generated samples is reported as the *category accuracy*.

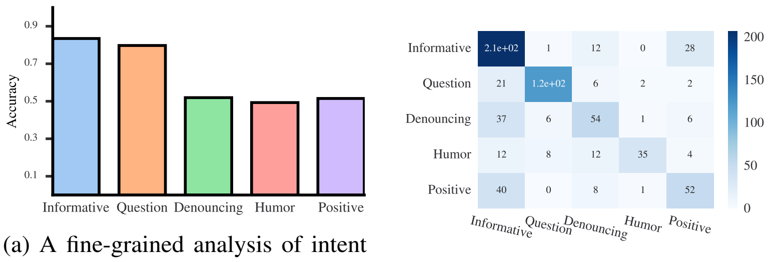
C Analysis of Intent-Conditioning

In order to systematically evaluate the effects of intent conditioning, we begin by analyzing the accuracy of the IC model for each intent separately. The results are depicted in Figure 5. From the bar chart, we observe that the accuracy of the intents – *informative* and *question*, is higher than the other intents, while *humor* displays the lowest accuracy. To obtain a more comprehensive understanding, the confusion matrix illustrates that the intents *denouncing* and *positive* tend to be recognized as *informative* by the IC model in some cases, while *humor* can also be recognized as *informative* and *denouncing*. Since the IC model is susceptible to errors, it is hard to say with certainty whether the generated counterspeech belongs to the desired intent, or whether the model has misclassified it. Hence, we utilize the confusion matrices from human evaluation and design a new metric in the next section for analyzing the intent conditioning due to the inherent reliability of human evaluators.

D Interpretability and Robustness of Intent Representations

A key advantage afforded by our approach is the exploration of interpretability, which is enabled by our paradigm of learning the intent representations separately. The intent representations illustrated in Figure 3 (*left*) depict that the intents *positive* and *denouncing* are both mapped closely to *informative*, and are slightly farther away from each other, while *question* and *humor* are considerably distant to all other intents. This observation is further supported by computing the cosine similarity in the original dimension of the representations (Fig. 7c). To assess the robustness of the obtained representations, we use implicit feedback from human evaluations to gauge the similarity between intents. We employ two strategies: (i) we design a new metric, *Implicit Similarity (IS)* to compute the similarity between pairs of intents implicitly through human evaluation responses without the knowledge of the actual intent; (ii) we utilize the intent information and use the confusion matrices obtained from human evaluation (Fig 6a) for this purpose.

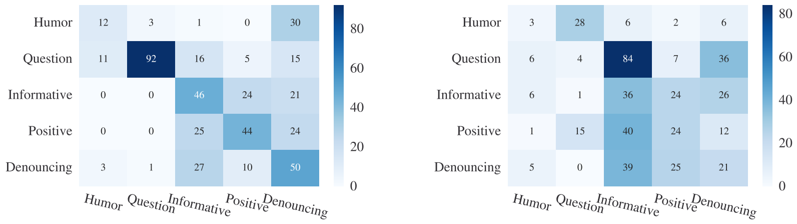
We plot the IS values for each intent pair in Figure 7. The IS scores for the pairs (I, D) and (I, P) are the highest, followed by the pair (P, D),



(a) A fine-grained analysis of intent identification accuracy of the generated outputs from QUARC on the test set as per the IC model.

(b) Confusion matrix depicting the intent classification (from the IC model) of the generated outputs from QUARC.

Figure 5: Automated evaluation of CA from the IC model for all intents. Note that *informative* and *question* achieve the highest accuracy demonstrating that QUARC is able to generate them more effectively than, say, humor, which achieves a relatively lower accuracy.



(a) Confusion matrix of the human evaluation for QUARC.

(b) Confusion matrix of the human evaluation for the best baseline, GPS.

Figure 6: Human evaluation heatmaps for QUARC and GPS. The rows represent the desired intent (the input given to the models) and the columns denote the intent labeled by human evaluators. Darker shade denotes a higher frequency of identification. For QUARC, all intents but *humor* are generally identifiable, while GPS is unable to condition on any intent effectively.

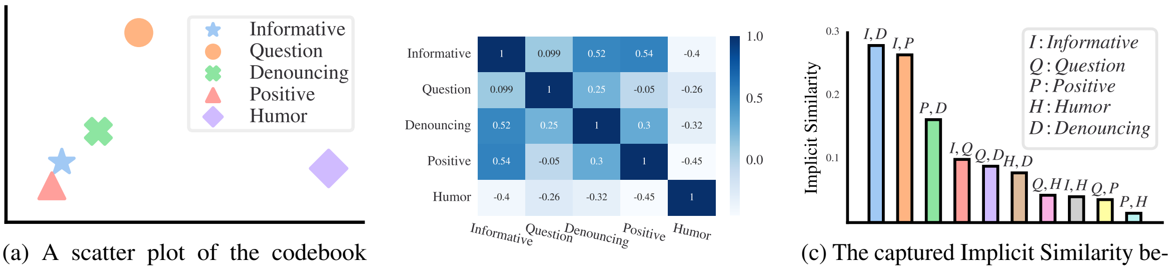
while the lowest scores are achieved by the pairs $(P, H), (Q, P), (I, H)$ and (Q, H). Interestingly, the IS scores *closely align* with the distances between the intent representations in the scatter-plot in Figure 3. This demonstrates the robustness of the intent representations learned by QUARC and highlights a critical factor responsible for its performance, as the representations align with the proximity that is inherently captured by evaluators.

Explicit Similarity through Human Evaluation: To further analyze the intent representations, we also utilize the desired intent c_i to generate the confusion matrices for human evaluation in Figure 6. We observe a similar pattern to that observed through IS, as we can see that the bottom-right 3×3 square has a darker shade as compared to the rest of the matrix, denoting that the *Informative, Positive* and *denouncing* intents are closer together when compared to other pairings.

E Human Evaluation

The evaluators recruited were well-versed in the field of NLP and social media. The form provided to them contained the descriptions of terminology such as *Hate Speech* and *Counterspeech*, and *Intents*. For further clarity, a few lines of description for each intent along with an example were also shown. The form also included information on the format of the questionnaire; the evaluators were made aware of how the evaluation data would be used in the study and were warned against the possibility of encountering foul or offensive language that could be upsetting.

Analysis: As shown in Figure 6, our model generates intent-identifiable outputs across all intents, with the exception of the *humor*, where the outputs were often assigned to *denouncing*. Conversely, GPS fails to effectively condition on intent, as evidenced by the mismatch between desired and obtained intents, with decent performance only on *informative*, perhaps due to its prevalence in the training set.



(a) A scatter plot of the codebook vectors corresponding to different intents after being reduced to a two-dimensional space through Principal Component Analysis (PCA).

(b) Visualization of the cosine similarity between the codebook vectors corresponding to different intents. Darker shade denotes higher similarity.

(c) The captured Implicit Similarity between all pair of intents. Note that (I, D) and (I, P) achieve the highest scores, while $(P, H), (Q, P), (I, H)$ and (Q, H) achieve the lowest scores.

Figure 7: Analysis and visualization of intent representations through: (a) dimensionality reduction to a 2-D space for plotting; (b) cosine similarity computed in the original dimension space of the representations. The similarity between *informative*, *positive* and *denouncing* is higher as compared to other intents. (c) The IS scores are closely aligned with the closeness of the representations in (a) and cosine similarities in (b). This serves to inform that the quantized representations learnt for each intent are demonstrably sound due to their similarity with human feedback.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7. Limitations

☑ A2. Did you discuss any potential risks of your work?
Section 8. Ethical Considerations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1. Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3. Dataset and Section 4. Methodology

☑ B1. Did you cite the creators of artifacts you used?
Section 3. Dataset. We extend the original dataset and cite the dataset.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 1: Introduction, Section 2: Related Works

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 1: Introduction

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3. Dataset

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 3

C ☑ Did you run computational experiments?

Section 5.Experimental Setup and Results

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B. Additional details on Experiments. Experimental setup

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B. Additional details on Experiments. Experimental setup

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix C: Further Analysis

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B. Additional details on Experiments. Experimental setup

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 5 : Experimental Setup and Results

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D: Human Evaluation Details

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 5 : Experimental Setup and Results

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 2: Related Works

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 5 : Experimental Setup and Results