Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI for Good

COUNTERGEDI: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech

Punyajoy Saha , **Kanishk Singh** , **Adarsh Kumar** , **Binny Mathew** and **Animesh Mukherjee**
Indian Institute of Technology, Kharagpur, India
{punyajoys,kanishksingh,adarshkumar712,binnymathew}@iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Recently, many studies have tried to create generation models to assist counter speakers by providing counterspeech suggestions for combating the explosive proliferation of online hate. However, since these suggestions are from a vanilla generation model, they might not include the appropriate properties required to counter a particular hate speech instance. In this paper, we propose COUNTERGEDI - an ensemble of generative discriminators (GEDI) to guide the generation of a DialoGPT model toward more polite, detoxified, and emotionally laden counterspeech. We generate counterspeech using three datasets and observe significant improvement across different attribute scores. The politeness and detoxification scores increased by around 15% and 6% respectively, while the emotion in the counterspeech increased by at least 10% across all the datasets. We also experiment with triple-attribute control and observe significant improvement over single attribute results when combining complementing attributes, e.g., *politeness*, *joyfulness* and *detoxification*. In all these experiments, the relevancy of the generated text does not deteriorate due to the application of these controls.

1 Introduction

One of the most effective strategies to combat the rising online hate speech is *counterspeech*. It is a direct response to hateful or harmful speech that seeks to undermine it. Several organisations like Facebook[1] have laid out guidelines for general public on how to counter hateful speech online. While these guidelines might be effective, writing a proper counterspeech is quite challenging [Fumagalli, 2020]. Recently, many non-profit organisations have taken up the task of countering online hate[2]. This nichesourcing of counterspeech can be effective but might be a mentally taxing task for the NGO operators given the amount of hate generated each day [Vidgen *et al.*, 2019]. To assist these operators, the scientific community have come up with different human-in-the-loop methods to collect counterspeech data [Fanton *et al.*, 2021] as well as build models to generate counterspeech suggestions [Zhu and Bhat, 2021].

Such generation models aim to reduce the human intervention by helping the counter speakers with suggestions[3], which they can further post-edit as per requirements. While these recent transformer based generation models can produce relevant outputs, they often fail to produce diverse output [Holtzman *et al.*, 2020]. Further, we cannot control the generated output for any attribute from the vanilla generation model. However, as pointed out by different authors [Bartlett and Krasodomski-Jones, 2015], counterspeech can vary based on the hate speech instance, demography of the hate and counter speakers [Mathew *et al.*, 2019] etc. Hence, the generation models without any control might produce suggestions that are not suitable for a particular instance. To counter effectively, the counter speakers should have the control over the generated content so that they can steer the counterspeech toward a desired property or a mixture of different properties. Such a generation model will allow the counter speaker to use different strategies to counter hate [Benesch *et al.*, 2016]. In this paper, we propose six different GEDI[4] models and use them in single and multiple attribute setting to guide the generation of a vanilla counterspeech generation model for English language. We observe that —

- The single attribute GEDI models improve the controlled attribute significantly for all the attributes. Specifically, there is 15% increase in politeness and 6% increase in detoxification.
- Among the GEDI with emotional attributes, *joy* has the highest attribute scores with an improvement of 15%, 17% and 88% in Gab [Qian *et al.*, 2019], Reddit [Qian *et al.*, 2019], and CONAN [Chung *et al.*, 2019] datasets respectively.
- While combining multiple attributes — *politeness*, *detoxification* and *joyfulness* (one of the emotional attribute) for controlling the generation, we observe a further increase in scores compared to the single attribute setup.
- We are able to maintain the relevancy of the generated

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

[2] https://wecounterhate.com/

[3] We view the counterspeech generating task as delicate task which requires human supervision.

[4] Generative discirimintators.

output measured using **BLEU-2** score while there is a slight drop in performance in terms of fluency which comes as a cost for the additional control.
- Finally, human judges also marked higher presence of the attributes in the attribute controlled setups compared to the baselines.

We show examples of emotion controlled counter speech corresponding to a hate speech in Table 1. Please refer to Appendix[5] for more examples on single and multi attribute controlled counterspeech. The code is added here[6].

Hate speech	*Anyone who isn't a re***ded pickle has privilege, I suppose.*
Counterspeech (no control)	*The R word is highly offensive, Spread the word to end the word!*
Counterspeech (sad)	*in the words of my favorite friend this post sounds like such hate speech. it makes you sound uneducated.* ***how sad****. do not use internet powers to silence your anger with something so simple or even reasonable!*
Counterspeech (joy)	*we all believe in open mindedness, whether your side of the fence is greater or lesser.* ***we hope you have fun!*** *mwahaha peace out and cool weather. keep hydrated npcs.*

Table 1: This table presents an example hate speech and the counterspeech generated using different control attributes which is marked in the (bracket). Please refer to the Appendix[5] for more examples on single and multi attribute controlled counterspeech.

2 Related Works

With regard to countering hatred, we focus on two research aspects relevant to the present work (i) counterspeech datasets and models for counterspeech generation (ii) controlling methods for generation models particularly in NLP.

Counterspeech generation strategies: An effective strategy to mitigate hate speech is counterspeech as it does not violate freedom of expression [Benesch *et al.*, 2016]. While the idea of countering some hateful messages is not new, the research community has recently started taking a massive interest in understanding counterspeech practices and its effectiveness while mitigating hate speech [Mathew *et al.*, 2019]. Recently, Tekiroğlu *et al.* proposed novel techniques to generate counterspeech using a GPT-2 model with post-facto editing by experts or annotator groups. One of the recent generation methods uses a three stage pipeline – *Generate, Prune* and *Select* (GPS) to generate diverse and relevant counterspeech output [Zhu and Bhat, 2021]. The primary challenge in counterspeech generation research is understanding if the counterspeech produced is effective. One study[Bartlett and Krasodomski-Jones, 2015] found that effectiveness of the counterspeech further depends on the tone of the counterspeech. In specific, sentimental or casual tone received 83% more responses [Frenett and Dow, 2015]. In addition to that, Mathew *et al.* found that different communities find different types of counterspeech effective. In this work, we present a novel approach to guide the counterspeech toward one or more desired properties. This is done by building a controllable generation pipeline based on GEDI.

Controllable text generation: Controllable text generation is the task of generating natural sentences whose attributes can be controlled. The previous approaches rely on reinforcement learning or training conditional generative models [Prabhumoye *et al.*, 2020]. One of the earliest line of work focused on controlling a desired attribute by side constraints [Sennrich *et al.*, 2016] and back propagating gradients [Dathathri *et al.*, 2019]. One of the variations - DEXPERTS [Liu *et al.*, 2021] uses language models for both positive and negative classes. While the other variation - GEDI [Krause *et al.*, 2020] uses class conditioned language models for positive and negative classes. The final output tokens in both these methods are generated based on an equation which utilises the contrast between the positive and the negative class.

We, in this work, use this GEDI model and apply it to the domain of counterspeech generation. We train different GEDI models to control attributes like *politeness*, *emotions* and *detoxification* of counterspeech.

3 Models

DialoGPT: We used a variant of the GPT model - DialoGPT [Zhang *et al.*, 2020] which was trained on a large corpus consisting of English Reddit dialogues. The corpus consist of 147 million instances of dialogues, collected over a period of 12 years. Unlike GPT-2, this model should generate better dialouge like responses to any given prompt. In this model, along with ground truth response $T = x_1, ..., x_n$ we also have a dialogue utterance history S. The model aims at maximising $p(T|S) = p(x_1|S) \prod_{i=2}^{n} p(x_i|S, ..., x_{i-1})$. For our experiment, we used DialoGPTm - a 24 layer, 345 million weight parameters transformer model[7] and finetune over a particular dataset having hate and counterspeech pairs.

Generate Prune Select (GPS): One of the recent counterspeech generation model is *Generate, Prune, Select* (*GPS*) – a three stage pipelined approach [Zhu and Bhat, 2021]. At first, the *generation part* generates a large number of diverse response candidates using a generative model based on RNN based autoencoder. Second, the *pruning part* prunes the ungrammatical candidates from the candidate pool. This is done using a classifier trained on linguistic acceptability classifier. Finally, the *response-selection part* selects appropriate responses based on the hate speech instance. We use the similarity based method *USE-LARGE-SIM* [Zhu and Bhat, 2021].

GEDI: For controlling generated counterspeech, we use a recent method Generative Discriminators (GEDI) [Krause *et al.*, 2020], where the authors present a decoding time algorithm to control the output from the generation model. GEDI assumes we have class conditioned language model (CC-LM) with a desired control code c and an undesired control code $\bar{c}$. For our case, we fix the control code c as 'true' and $\bar{c}$ as 'false'. For each dataset, the attribute mentioned in the $+ve$ column in Table 3 is considered as desired, while $-ve$ column is used as undesired attribute.

The authors use the contrast between $P_\theta(x_{1:t}|c)$ and $P_\theta(x_{1:t}|\bar{c})$ to guide sampling from an LM that gives

[5] https://tinyurl.com/bdcmk9nm

[6] https://github.com/hate-alert/CounterGEDI

[7] https://huggingface.co/microsoft/DialoGPT-medium

$P_{LM}(x_{1:T})$. The probability that the next token x_t belongs to desired class is calculated using this contrast.

For controlled generation, the authors propose a simple method to guide the model toward the target class which is represented using the heuristic equation 1 where ω is controllable parameter. In order to control multiple attributes, we extend the heuristic as represented in equation 2 where ω_i is a controllable parameter to bias the generation toward class c_i for GEDI trained on the i^{th} attribute.

$$P_w\left(x_t \mid x_{<t}, c\right) \propto P_{LM}\left(x_t \mid x_{<t}\right) P_\theta\left(c \mid x_t, x_{<t}\right)^\omega \tag{1}$$

$$P_w\left(x_t \mid x_{<t}, c_1, ..c_n\right) \propto P_{LM}\left(x_t \mid x_{<t}\right) \prod_{i=1,..,n} P_\theta\left(c_i \mid x_t, x_{<t}\right)^{\omega_i} \tag{2}$$

4 Datasets

Counterspeech datasets: In order to evaluate our approach we use three public datasets which contain hate speech and its corresponding counterspeech. The details of these datasets are noted in Table 2. Reddit and Gab datasets contain $5,257$ and $14,614$ hate speech instances respectively [Qian *et al.*, 2019]. We use the English part of the CONAN dataset [Chung *et al.*, 2019] which contains 408 hate speech instances. The counterspeech in Gab and Reddit datasets were written by AMT workers, whereas for CONAN the counterspeech was written by expert NGO operators.

We further made hate speech and counterspeech pairs from these datasets such that each hate speech was associated with one counterspeech. Finally, we ended up with 3,864, 14,223, 41,580 datapoints for CONAN, Reddit and Gab respectively. We split each dataset randomly into train, validation, test set with 80% for training, 10% each for validation and testing.

Dataset	Source-H	Source-C	Hate instances	Total pairs
CONAN	synthetic	expert	408	3,864
Reddit	reddit	crowd	5,257	14,223
Gab	gab	crowd	14,614	41,580

Table 2: This table presents the source of hate speech (Source-H), source of counterspeech (Source-C), hate speech instances and the total pairs for each of the CONAN, Reddit and Gab dataset.

Attribute datasets: We control several attributes in the generated counterspeech. We selected these attributes following the recommended strategies for counterspeech [Benesch *et al.*, 2016] and properties of responses in human conversation [Clark *et al.*, 2019].

Politeness: One of the properties of counterspeech as suggested by [Benesch *et al.*, 2016] is empathy. As a first step in that direction we tried to make the generated counterspeech more polite. We used the dataset of 1.39 million posts released by [Madaan *et al.*, 2020] labelled into nine politeness classes (P1-P9). As recommended by the authors, we considered P9 as the polite part and others (P1-P8) as non-polite.

Detoxification: [Benesch *et al.*, 2016] also noted several strategies which are discouraged while writing counterspeech. One of these discouraged strategies are hostile or aggressive behaviour. To detox any hostile counterspeech generated by the generation model, we use the a popular Kaggle dataset[8] which contains text samples having 'toxic' and 'non-toxic' labels. We stratified-split the released training dataset randomly into 90% training and 10% validation sets. The test set is already released separately with the dataset. We trained a GEDI model considering toxic as the positive label and non-toxic as the negative label. While generating using GEDI, we guide the generation toward the negative class (non-toxic).

Emotion: Another important aspect of conversation is communicating different emotions. A study [Prendinger and Ishizuka, 2005] found that systems expressing emotions are more capable of providing user satisfaction. In case of counterspeech, emotions might enhance the effect of the generated counterspeech [Benesch *et al.*, 2016]. For example, 'sadness' as an emotion can be added when the counter speakers affiliate themselves with the target group. Similarly, 'joy' can used to convey positivity in the counterspeech.

In order to control the emotion while generating a counterspeech, we used a large dataset [Saravia *et al.*, 2018] of 416,809 datapoints comprising posts having seven emotions – 'sadness', 'joy', 'fear', 'anger', 'surprise', and 'love'. For this paper, we did not consider - 'love' and 'surprise' emotions as these had less than 10% posts in the dataset. We stratified-split each dataset randomly into training, validation, and test set with 80% for training, and 10% for both validation and testing. We consider each emotion as a separate attribute and trained a GEDI model for that emotion by considering it as positive label and other emotions as negative labels. For our experiments, we primarily focus on guiding the models toward the positive class.

A summary statistic of the attribute dataset for each of the task considered is noted in Table 3.

Dataset	+ve	-ve	T_r (%+ve)	V (%+ve)	T_e (%+ve)
Polite	p	n-p	1.12M (20%)	137k (20%)	137k (20%)
Toxic	t	n-t	143k (10%)	16k (10%)	153k (4%)
Emotion	j	o	333k (34%)	42k (34%)	42k (34%)
	f	o	333k (11%)	42k (11%)	42k (11%)
	s	o	333k (29%)	42k (29%)	42k (29%)
	a	o	333k (14%)	42k (14%)	42k (14%)

Table 3: This table shows the attribute datasets, positive and negative classes and data present in train, validation and test part for each. T_r: Train, V: Validation, T_e: Test, p: polite, n-p: non-polite, t: toxic, n-t: non-toxic, s: sadness, j: joy, a: anger, f: fear, o: others. The % associated with the T_r, V and T_e are the % of positive labels.

5 Experimental Setup

Counterspeech generation models: The DialoGPTm model for each counterspeech model has six initial layers fixed due to resource constraints. The model were trained till 10 epochs with batch size as 8. We saved the final model at the epoch having the best language modelling loss for the validation dataset. We used a maximum length of 256 tokens for the DialoGPTm model[9]. The learning rate is fixed at $5e^{-6}$.

GEDI models: We train the GPT-2 model as the GEDI models based on the training setting specified in the original pa-

[8] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[9] 1% datapoints have more than 256 tokens.

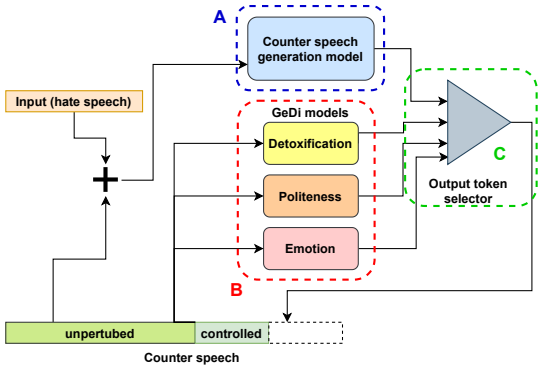

Figure 1: The figure shows how the overall setup of the pipeline. The counterspeech generation model is produces a probability distribution for the next possible token using hate speech+ counterspeech generated till now.

per [Krause *et al.*, 2020]. For each model we fix the batch size at 8 and train the models for 5 epochs. The λ weight in the loss equation is fixed at 0.8 to maximise generation quality for the GEDI model. We used a maximum length of 128 tokens for the GPT-2 model. The learning rate is fixed at $2e^{-5}$ for training the model following the recommendations by [Krause *et al.*, 2020].

Final pipeline: Our final pipeline comprises three parts as shown in Figure 1. The **part A** represents the vanilla counterspeech generation model trained on one of the three counterspeech datasets. Similar to an auto-regressive setup, it takes in the hate speech with the currently generated counterspeech (empty at the initial step) and produces next token probablities for the production of the next token. The **part B** consists of the single or multiple GEDI models. Each GEDI model controls an attribute out of the total six. It takes as input the currently generated counterspeech and produces token probabilities based on the desired attribute. We initially allow the counterspeech generation models to generate 10 tokens without any control to provide the initial prompt to the GEDI model. Finally, **part C** selects the next token based on the token probabilities from different models, i.e., the counterspeech generation model and the GEDI models following equation 2. For each GEDI model, we primarily control the weight (ω) as mentioned in equation 1, while other parameters are kept same as the paper [Krause *et al.*, 2020]. For single attribute, we fix the weight at 1 to give equal importance to the counterspeech generation as well as the control attributes. For two attribute control, we set weights at 0.5 for both the attributes. For three attributes control, which comprises detoxification, politeness and an emotion, we set 0.3 for politeness & detoxification each and 0.4 for the emotion. We also use nucleus sampling as a decoding strategy [Holtzman *et al.*, 2020]. Please check Appendix [5] for more details.

6 Evaluation

We consider several metrics to evaluate our whole pipeline of controlled counterspeech generation. The *generation metrics* measure the generation capability of the DialoGPTm and GEDI models. The *classification metrics* are mainly to evaluate the GEDI model on the attribute datasets. Finally, we measure the amount of control in the generated counterspeech using external classifiers which we refer to as *controller metrics*. We generate 5 samples for every hate speech instance with DialoGPTm. The GPS framework automatically selects the best response based on the heuristic, hence we keep one sample for every hate speech instance.

Generation metrics: To measure the generation quality, we use different standard metrics. We use *BLEU-2*[10] and *METEOR* [Sai *et al.*, 2020] to measure how similar the generated counterspeech are to the ground truth counterspeech. We also measure if the generation model generates a diverse and novel counterspeech using metrics from previous research [Wang and Wan, 2018]. To measure fluency, we use a classifier of linguistic acceptability trained on the COLA dataset [Warstadt and Bowman, 2019]

GEDI metrics: For classification, we report *accuracy*, *macro F1-score*, and *AUROC* score for each GEDI model's performance on a test dataset of a particular attribute. We also report the generation performance using the perplexity [Zhang *et al.*, 2020].

Controller metrics: In order to evaluate the ability of the GEDI controller to control the attribute, we used third-party classifiers for each attribute. For politeness, we trained a bert-base-uncased model for politeness level detection on a scale of 0 to 7[11]. For measuring emotion in the generated text, we used the Ekman version of the GoEmotions models[12]. For each post, it returns a confidence score between 0-1 for anger, disgust, fear, joy, sadness, surprise + neutral. We report the confidence score for a particular emotion as a measure of that emotion in a given post. Finally, to measure toxicity we used the HateXplain model [Mathew *et al.*, 2021] trained on two classes – toxic and non-toxic[13]. We report the confidence between 0-1 for the non-toxic class.

7 Results

Generation results: We compare DialoGPTm model with the GPS in Table 4. We find **BLEU-2** scores are better for the GPS model while the **METEOR** scores are better for DialoGPT model for all three datasets. DialoGPTm is also better in terms of novelty and diversity for all the three datasets. The fluency metric **COLA** is better for GPS for all the three datasets, since it straightforwardly prunes grammatically incorrect samples. Since DialoGPTm presents a competitive performance compared to the state-of-the-art model, we therefore use DialoGPTm for the rest of the experiments.

GEDI metrics: As reported in Table 5, we find that F1-score and AUCROC scores for politeness and all the four emotions are above 0.9. This highlights that even with 0.2 as the weight for the discriminator we are able to get good scores on classification. The perplexity scores for all the test datasets are

[10] converted to a scale of 0-100 from 0-1

[11] https://github.com/AlafateABULIMITI/politeness-detection

[12] https://huggingface.co/monologg/bert-base-cased-goemotions-ekman

[13] https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two

Model	B2 (↑)	COLA (↑)	M (↑)	N (↑)	D (↑)
CONAN					
GPS	**41.5**	**0.82**	0.14	0.18	0.60
DialoGPTm	12.7	0.78	**0.18**	**0.84**	**0.80**
Reddit					
GPS	**14.1**	**0.82**	0.11	0.30	0.47
DialoGPTm	6.9	0.75	**0.17**	**0.82**	**0.74**
Gab					
GPS	**13.9**	**0.82**	0.12	0.15	0.41
DialoGPTm	7.7	0.80	**0.17**	**0.80**	**0.72**

Table 4: Evaluation results for the three datasets. We report BLEU-2 (B2), COLA, METEOR (M), novelty (N) and diversity (D) to compare the two baselines: generate-prune-select (GPS) framework and DialoGPTm. For all metrics, higher is better and **bold** denotes the best scores.

also around 3.5[14]. GEDI model for toxicity has lower scores than the other attribute tasks. The F1-score for toxicity detection is ~ 0.6 and AUCROC is ~ 0.83. The perplexity is also higher at around 4.5 for the toxicity dataset. This highlights the difficulty of the task of detecting toxicity.

Dataset	**Positive**	**F1** (↑)	**Acc** (↑)	**AUC**(↑)	**Perplexity** (↓)
Toxicity	toxic	0.60	0.85	0.84	4.428
Politeness	polite	0.93	0.96	0.93	3.476
Emotion	joy	0.96	0.96	0.97	3.546
Emotion	sadness	0.98	0.98	0.99	3.543
Emotion	fear	0.94	0.97	0.98	3.774
Emotion	anger	0.96	0.98	0.99	3.560

Table 5: GEDI generation and classification performance on test set of attribute datasets. Generation is evaluated using the Perplexity whereas classification performance is measured using F1-score (F1), Accuracy (Acc) and AUCROC (AUC). For all the metrics except perplexity, higher is better.

Single-attribute control: In Table 6, we report the amount of different attributes present in the generated counterspeech for each dataset and for each model. When we compare GPS and DialoGPTm, we find that except anger emotion, all other scores are significantly higher for DialoGPTm. Second, using control for a particular attribute significantly improves the presence of that attribute (p-value <0.001). For instance, in Table 6, the politeness score increases from 3.91 to 4.54, from 5.24 to 6.05 and 5.14 to 6.11 for CONAN, Reddit and Gab respectively when the DialoGPTm model is controlled for politeness. This is true for all attributes barring the 'anger' emotion. Politeness and detoxification score increased by 15-18% and 6-8% respectively across all the datasets. For the emotion attributes, 'joy' has the highest scores among all for both controlled and uncontrolled attribute. We see an overall increase in 'joy' of around 17% for Gab, 14% for Reddit and 88% for CONAN. Counter responses in CONAN datasets are mostly devoid of any emotions hence bringing a change in them is much easier than the Reddit/Gab datasets which are higher in terms of the joy attribute. We reach closer to GPS baseline for anger emotion while controlling anger emotion and increase the score by 54%, 55% and 16% for Reddit, Gab and CONAN, respectively. While the increase for other emotions – 'sadness' and 'fear' increased significantly, the overall scores for them remain low.

[14]For reference, perplexity for pretraining GPT-2 comes around 10 after 10K steps (https://tinyurl.com/3vwrvscd).

Model	**D** (↑)	**P** (↑)	**J** (↑)	**A** (↑)	**S** (↑)	**F** (↑)
CONAN						
GPS	**0.68**	2.01	0.16	**0.12**	0.03	0.01
DialoGPTm	0.64	3.91	0.18	0.09	0.04	0.01
DialoGPTm-c	**0.68**	**4.54**	**0.34**	0.11	**0.08**	**0.05**
Reddit						
GPS	0.82	1.62	0.23	**0.32**	0.04	0.01
DialoGPTm	0.82	5.24	0.63	0.17	0.06	0.00
DialoGPTm-c	**0.87**	**6.05**	**0.72**	0.27	**0.10**	**0.02**
Gab						
GPS	0.79	1.46	0.22	**0.28**	0.04	0.01
DialoGPTm	0.81	5.14	0.66	0.17	0.05	0.00
DialoGPTm-c	**0.85**	**6.11**	**0.77**	0.26	**0.10**	**0.02**

Table 6: Performance of single attribute setups with the vanilla baseline generate-prune-select (GPS) and DialoGPTm models. The attributes measured according to column are politeness (P), detoxification (D), sadness (S), joy (J), anger (A) and fear (F). Politeness (P) is measured in a scale of 0-7 whereas others are measured in the scale $[0, 1]$. For the last row - controlled DialoGPTm (DialoGPTm-c) the column name also represents the attribute getting controlled. For all the metrics, higher is better and **bold** denotes the best scores.

Scores	**Detox**	**Polite**	**Joy**	**Anger**	**Sadness**	**Fear**
CONAN						
BLEU-2	**13.8**	12.1	12.2	11.6	12.0	12.8
COLA	**0.83**	0.72	0.72	0.74	0.76	**0.72**
Reddit						
BLEU-2	**8.1**	7.8	7.7	7.8	7.5	7.3
COLA	0.72	0.77	0.70	0.72	**0.81**	0.70
Gab						
BLEU-2	**8.7**	8.3	8.5	8.3	8.2	8.3
COLA	**0.85**	0.82	0.76	0.76	0.80	0.78

Table 7: BLEU-2 and COLA performance for single attribute setups for DialoGPTm-c model. Each column name represents the individual attribute model namely politeness (P), detoxification (D), sadness (S), joy (J), anger (A) and fear (F). **Bold** denotes the best scores across the row.

Multi-attribute control: We also generate counterspeech with the DialoGPTm with mutli-attribute control. We keep politeness, detoxification and one of the emotion[15] as control attributes. This gives us four variations for each dataset. We then measure the individual attribute scores for each of these three attribute and report the results in Table 8. For detoxification scores, the setup - $joy{+}polite{+}detox$ outperforms other setups across all the experiment. This setup even outperforms the single-attribute detoxification setup by 8%, 2% and 2% for CONAN, Reddit and Gab, respectively. For politeness score, the best performance occurs for $joy + polite + detox$ setup for CONAN and Reddit dataset, while the setup - $fear + polite + detox$ performs better in case of the Gab dataset. Compared to single attribute setup for politeness, the politeness scores drop across all the multi-attribute setups. Among the emotions, the attribute score for 'joy' in a multi-attribute setting outperforms the single attribute setting by 44%, 13% and 10% for CONAN, Reddit and Gab. For 'anger', the scores in multi-attribute setting decrease around 25-30% when compared to the single attribute setting. For other attributes like 'sadness' and 'fear', the multi-attribute

[15]One among 'joy', 'anger', 'fear' and 'sad'.

results are below 0.1, similar to the single attribute results. Please also see Appendix [5] for ablation studies.

Quality of controlled generation: In the previous section, we observed that we were able to control attributes in generated outputs in single and multi-attribute setups. While this is encouraging, it is important to understand if the controlled text are losing the central theme of remaining a counterspeech and are still fluent. For the former, we measure the **BLEU-2** metric and for the latter we use the **COLA** metric.

According to Table 7, we find that relevance of the output (measured using BLEU-2) does not change much across different attributes for the single attribute setups. For some of the attributes like detoxification, the BLEU-2 scores even outperform the DialoGPTm model (without control) for all the datasets as noted in column B2 in Table 4. For Reddit and Gab, there is a further improvement of 1-2 points in the **BLEU-2** metric for other attributes also as compared to the vanilla DialogGPTm model (in column B2 in Table 4). This shows that the controls do not affect the overall relevance of the generated counterspeech. In fact, the relevance improves in many cases. In terms of fluency, we see a slight drop which comes as a cost for controlling different attributes except few cases (comparing column COLA in Table 4 and Table 7). This might be due to the fact that GEDI model is not geared toward maintaining the fluency of the models. The observation holds for the multi attribute setup as well (comparing columns B2 and COLA in Table 4 and Table 8).

Overall, we observe that it is possible to control the attributes in the generated outputs using the single attributes. Our experiments with multi-attributes further reveals that there are certain complementing attributes for e.g $joy + polite + detox$ which can be used to further increase the single-attributes setups. For other setups, the attribute scores drops below the single attribute setups. Another promising observation is that the control of attribute does not harm the relevance of the generated output as they still remain close to the ground truth. Since GEDI is not geared toward improving the fluency, we see a slight drop in the fluency of the generated outputs. An interesting research direction would be to look into improving attribute and fluency scores while using multi-attribute setups. We have added examples of single and multi-attribute setup in the Appendix [5].

Human evaluation: In order to understand, if the improvement in the attribute scores across (while controlling different attributes) would be visible to the moderators, we perform a human evaluation on the generated counterspeech. In this experiment, an annotator is shown three sentences - one generated from the GPS pipeline, another generated using DialoGPTm model and finally, one generated using the DialoGPTm model where some attribute x was getting controlled. We hide the type of model from which the post was generated and further shuffle the posts to remove any ordering bias. Next, the annotator was asked to mark the amount of the attribute x in the given three posts on a scale of 0-5 where 0 presents the absence of the attribute while 5 corresponds to the highest presence of that attribute. Five annotators participated in the annotation with each post getting marked by two annotators. The annotators annotated 20 randomly selected triplets per dataset for each attribute. We do not include the detoxification attribute for these experiments as there is very little difference in detoxification scores when comparing the baseline and the controlled setups. For more details about the annotations, please refer to the Appendix[5].

Attributes	Detox(↑)	Polite(↑)	Emotion(↑)	B2(↑)	COLA(↑)
CONAN					
Joy(J)+P+D	**0.74**	**4.13**	0.49 (J)	13.4	**0.79**
Anger(A)+P+D	0.67	3.06	0.08 (A)	12.6	0.68
Sad(S)+P+D	0.70	3.56	0.07 (S)	13.2	0.74
Fear(F)+P+D	0.70	4.00	0.06 (F)	**13.6**	0.75
Reddit					
Joy+P+D	**0.89**	**5.79**	0.82 (J)	8.3	**0.81**
Anger+P+D	0.85	4.24	0.19 (A)	**8.3**	0.72
Sad+P+D	0.87	3.56	0.09 (S)	8.2	0.79
Fear+P+D	0.87	4.00	0.01 (F)	7.8	0.79
Gab					
Joy+P+D	**0.87**	5.68	0.85 (J)	**8.8**	**0.85**
Anger+P+D	0.83	4.11	0.19 (A)	8.5	0.75
Sad+P+D	0.85	4.70	0.09 (S)	**8.8**	0.84
Fear+P+D	0.86	**5.82**	0.01 (F)	**8.8**	0.83

Table 8: Results of controlling three attributes – politeness, detoxification and one of the emotions in a multi-attribute setting. The columns represent the amount of the attribute present for each setup. The column – *emotion* represents the score of the emotion shown in the parenthesis that is being controlled for that instance. BLEU(B2) and COLA were also reported for different setups. For all metrics, higher is better and **bold** denotes the best scores.

Model	Polite (↑)	Joy (↑)	Anger (↑)	Sad (↑)	Fear (↑)
CONAN					
GPS	0.50	1.30	2.50	**1.00**	0.00
DGPTm	0.59	**2.50**	3.00	0.75	0.75
DGPTm-c	**2.00**	1.00	**4.00**	**1.00**	**2.00**
Reddit					
GPS	1.83	0.93	1.50	0.33	0.36
DGPTm	2.66	2.50	1.50	0.66	**1.33**
DGPTm-c	**3.50**	**3.33**	**2.00**	**2.00**	1.25
Gab					
GPS	1.56	1.28	0.81	0.4	0.17
DGPTm	2.17	2.50	1.66	1.11	0.89
DGPTm-c	**3.21**	**2.92**	**1.90**	**2.03**	**1.00**

Table 9: Average human judgement scores (scale 0-5) for each of the models – GPS, DialoGPTm and controlled DialoGPTm (DGPTm). Each column represents the attribute that DialoGPTm-c (DGPTm-c) is controlled for. For all the metrics, higher is better and **bold** indicates best scores.

We observe an improvement in most of the attribute scores for the controlled model over the two baselines. Three cases where the improvement is not present is while controlling 'joy' and 'sad' for the CONAN dataset and controlling 'fear' for Reddit dataset. While controlling attribute 'sad', we only see an improvement relative to the base DialoGPTm model. The summary of this experiment is presented in the Table 9.

8 Conclusion

Our research aims to add controllable parameters to counterspeech generation setup which can help the moderators to tune the counterspeech toward a particular strategy. Our controllable GEDI models for six different attribute shows significant improvement in the attribute scores over the baselines. We also try to control the generation using multi-attribute and find that the attribute scores can increase further if suitable attributes are mixed together.

Ethical Statement

Hate speech is a complex phenomenon. While the language generation methods are better than before, it is still very far from generating coherent and meaningful replies [Bender *et al.*, 2021]. Hence, we advocate against deployment of fully automatic pipelines for countering hate speech [de los Riscos and D'Haro, 2021]. Based on the current progress, in this pipeline, an active participation of the counter speakers is required to generate relevant counterspeech. This automation, in turn, has the potential to reduce the mental toll of the counter speakers, at least partially.

References

[Bartlett and Krasodomski-Jones, 2015] Jamie Bartlett and Alex Krasodomski-Jones. Counter-speech examining content that challenges extremism online. *DEMOS, October*, 2015.

[Bender *et al.*, 2021] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, 2021.

[Benesch *et al.*, 2016] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. Considerations for successful counterspeech. *A report for Public Safety Canada under the Kanishka Project. Accessed November*, 25:2020, 2016.

[Chung *et al.*, 2019] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*, 2019.

[Clark *et al.*, 2019] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. *What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents*. ACM, 2019.

[Dathathri *et al.*, 2019] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *ICLP*, 2019.

[de los Riscos and D'Haro, 2021] Agustín Manuel de los Riscos and Luis Fernando D'Haro. *ToxicBot: A Conversational Agent to Fight Online Hate Speech*. 2021.

[Fanton *et al.*, 2021] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *NAACL*, 2021.

[Frenett and Dow, 2015] Ross Frenett and Moli Dow. One to one online interventions: A pilot cve methodology. *Institute for Strategic Dialogue*, 2015.

[Fumagalli, 2020] Corrado Fumagalli. Counterspeech and ordinary citizens: How? when? *Political Theory*, page 0090591720984724, 2020.

[Holtzman *et al.*, 2020] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLP*, 2020.

[Krause *et al.*, 2020] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*, 2020.

[Liu *et al.*, 2021] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *ACL-IJCNLP*, 2021.

[Madaan *et al.*, 2020] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. In *ACL*, 2020.

[Mathew *et al.*, 2019] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *ICWSM*, 2019.

[Mathew *et al.*, 2021] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, 2021.

[Prabhumoye *et al.*, 2020] Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*, 2020.

[Prendinger and Ishizuka, 2005] Helmut Prendinger and Mitsuru Ishizuka. The empathic companion: A character-based interface that addresses users' affective states. *App. AI*, 2005.

[Qian *et al.*, 2019] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In *EMNLP-IJCNLP*, 2019.

[Sai *et al.*, 2020] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *arXiv preprint arXiv:2008.12009*, 2020.

[Saravia *et al.*, 2018] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *EMNLP*, 2018.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *NAACL 2016*, 2016.

[Tekiroğlu *et al.*, 2020] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. Generating counter narratives against online hate speech: Data and strategies. In *ACL*, 2020.

[Vidgen *et al.*, 2019] Bertie Vidgen, Helen Margetts, and Alex Harris. How much online abuse is there. *Alan Turing Institute*, 2019.

[Wang and Wan, 2018] Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, 2018.

[Warstadt and Bowman, 2019] Alex Warstadt and Samuel R Bowman. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*, 2019.

[Zhang *et al.*, 2020] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL: System Demonstrations*, 2020.

[Zhu and Bhat, 2021] Wanzheng Zhu and Suma Bhat. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625*, 2021.