CounterGeDi : A controllable approach to generate polite, detoxified and emotional counterspeech

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee

Work accepted at IJCAI-ECAI 2022

⚠️ *This presentation contains material that is **offensive** or **hateful**; however this cannot be avoided owing to the nature of the work.*

Hate-Speech

Hate-Speech is a language used to express hatred toward a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation etc.

Ways to Combat Hate-Speech

Inaction : By not responding to the hate-speech.
Deletion : Deleting or Suspending the user account is the most common way used by online platforms such as Facebook and Twitter.
CounterSpeech : Directly intervening with textual response that counter the hate-content.

What is CounterSpeech?

"*Given an instance of hate-speech, intervene directly into the conversation through automated fact-bound or empathetic arguments that can effectively neutralize hate-speech and incite constructive dialogue.*"

Why CounterSpeech?

- Suspension/Removal of posts is a threat to doctrine of free speech.
- Efficient way even with the most virulent and aggressive voices.
- Semi-automated counter-narration is more scalable, sensitive and capable of fighting extremism..

Related works

Counter speech generation is mostly focused with generating more relevant counter speech -

- Fine-tuning pretrained generation models -GPT2, DialogGPT [1]
- Additional pruning or selection pipelines [2]

[1] Pranesh, Raj Ratn, Ambesh Shekhar, and Anish Kumar. "Towards automatic online hate speech intervention generation using pretrained language model." (2021).
[2] Zhu, Wanzheng, and Suma Bhat. "Generate, Prune, Select: A Pipeline for Counterspeech Generation\\against Online Hate Speech." *Findings of the Association for Computational Linguistics* (2021).

Related works

Counter speech generation is mostly focused with generating more relevant counter speech

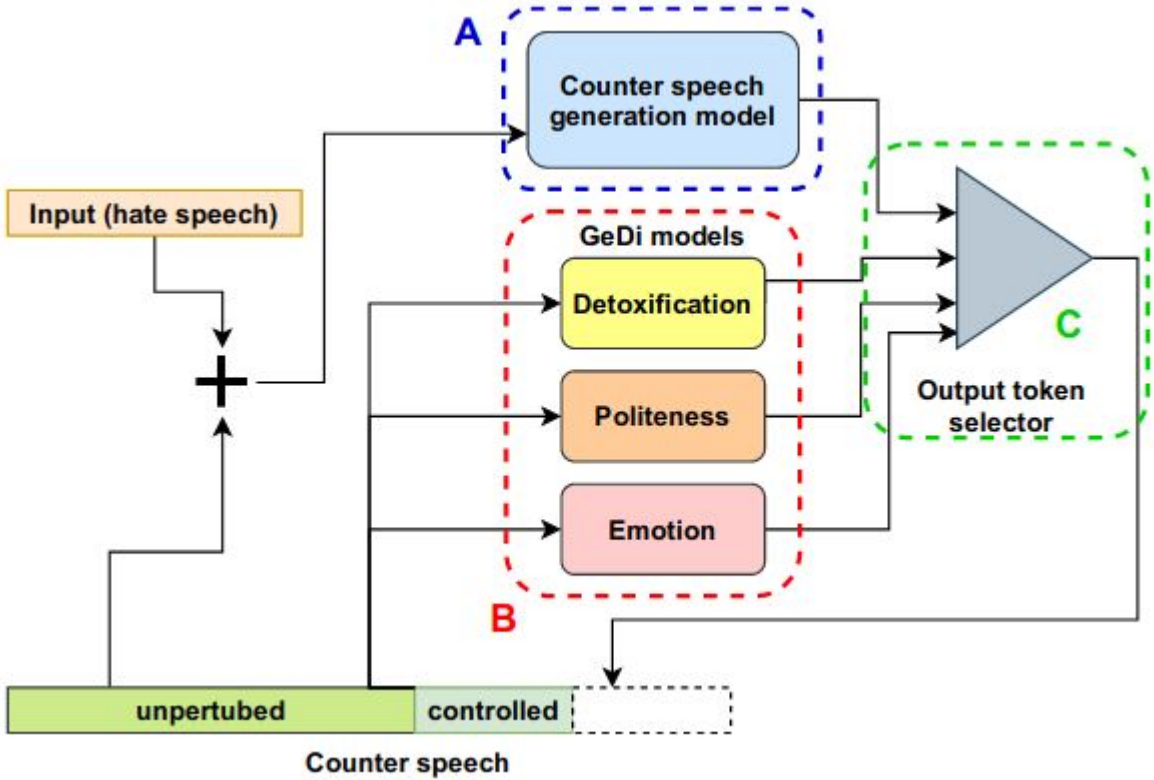
But, tone is also important -

- Sentimental or casual tone received 83% more response [1]
- Mathew et al. [2] found that different communities find different types of counterspeech effective .
- Empathy based counter speech can help reduce the racist comments [3]

[1] Ross Frenett and Moli Dow. One to one online interventions: A pilot cve methodology. Institute for Strategic Dialogue, 2015.
[2] 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.
[3] Hangartner, Dominik, et al. "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment." Proceedings of the National Academy of Sciences 118.50 (2021).

Can we add additional control to make the tone of counter speech better ?

Our proposal



Baseline Models

- A conversation sequence can be represented as turn of dialogues as U_t and $R_t = \{u_1, r_1, \dots, u_t, r_t\}$ where u_t is the user input and d_t is the response.
- Upon concatenation of U_t and R_t, $X_t = \{x_1, x_2, \dots, x_{i-1}\}$ is the sequence of tokens
- By chain rule of probability, the language model distribution can be computed as:
  - $p(X) = \prod_{i=1} p(x_i \mid x_0, x_1, \dots, x_{i-1})$
- Here we use a transformer based language model namely **GPT-2**[1] and **Dialo-GPT**[2].

[1] https://huggingface.co/gpt2
[2] https://github.com/microsoft/DialoGPT

Couterspeech Datasets

Dataset	Labels	Total size	Language	Source	Target Community
Qian et al., '19	Hate-Intervention Pairs	3,847	English	**REDDIT**	**Mixed**
Qian et al., 19	Hate-Intervention Pairs	11,169	English	**GAB**	**Mixed**
Chung et al. '19	Hate-Intervention Pairs	408	English, Italian French	**CONAN**	**Muslims**

Note:- None of these dataset have additional labels to control the tone of the counter speech by supervision

Controllable Text Generation

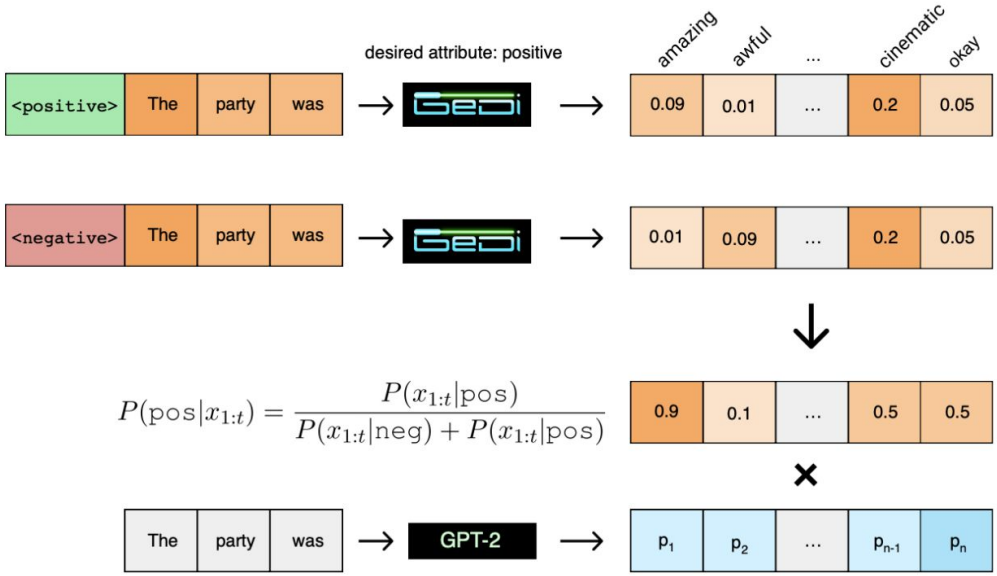
- We steer the generation model to contain certain quality attributes such as:
  - **Emotion** - Generating more diverse responses catering to large number of communities.
    - Sadness - Show affiliation with the targeted communities.
    - Joy - Convey positivity in the counterspeech.
    - Anger - Express disagreement with the speaker.
  - **Politeness** - Toward more empathetic counterspeech.
  - **Detoxification** - Minimize hostile behaviour (slur words) in generated responses.

Attribute Datasets

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.

GEDI: Generative Discriminator Guided Sequence Generation



Attribute-Control

Trained separate GEDI Models for each Controllable Parameter

Single-Attribute Control

- Combined single attribute GEDI Model with Fine-tuned base Dialo-GPT model.
- Equi-weighted Combination of Probabilities at the time of generation.

Multi-attribute Control

- Combined of several single attribute GEDI Models with Fine-tuned base Dialo-GPT model.
- Equal Weights provided for each attribute while combining probabilities at the time of generation
- Example: Polite + Detox + [Emotion] + Base model

Multi-attribute Sampling

- **Generative Loss**

$$\mathcal{L}_g = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log P_\theta\left(x_t^{(i)} \mid x_{<t}^{(i)}, c^{(i)}\right)$$

- **Discriminative Loss**

$$\mathcal{L}_d = -\frac{1}{N}\sum_{i=1}^{N}\log P_\theta\left(c^{(i)} \mid x_{1:T_i}^{(i)}\right)$$

- **Weighted-Loss**

$$\mathcal{L}_{gd} = \lambda\mathcal{L}_g + (1-\lambda)\mathcal{L}_d$$

- **Single-Attribute Sampling**

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

- **Multi-Attribute Sampling**

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

Experimental Setup

Models considered for Experiments:

- **Baseline 1: Generate, Prune, Select (GPS)**
  - A three stage pipelined approach for counterspeech generation, Zhu and Bhat [2021]

- **Baseline 2: Dialo-GPT Fine-tuned Base model**
  - Used a variant of the GPT model - Dialo-GPT Zhang et al. [2020]
  - Fine-tuned on respective datasets: CONAN, Reddit and Gab

- **CounterGEDI: Single Attribute and Multi-Attribute**
  - Our model with GEDI models trained for Different controlling attributes
  - Generation performed with single and Multi-attribute combination

Metrics

- **Generation Metrics**
  - Novelty, Diversity, BLEU (relevancy) and COLA (Fluency)
- **GEDI-Metrics**
  - Generation and classification performance on test-set of attribute datasets.
- **Controller Metrics :** We use third-party classifiers for evaluating each attribute.
  - **Politeness :** Trained a bert-base-uncased model for politeness detection[1].
  - **Emotion :** Used the Ekman version of Go-Emotions Model[2] .
  - **Detoxification :** Evaluated using HateXplain Model[3] 's confidence for the toxic class.

[1] https://github.com/AlafateABULIMITI/politeness-detection
[2] https://huggingface.co/monologg/bert-base-cased-goemotions-ekman
[3] https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two

Automatic evaluation

Baselines

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.

Fine-tuned Results on REDDIT/GAB

Model	Text
(Instance)	No, because she's literally a retard and hasn't come to terms with that fact just yet.
GPS	Its inappropriate to use words that directly attack one based on their mental health (retard) and sexual orientation (homos).
Dialo-GPT	Please refrain from using hateful ableist language in your posts.
(Instance)	Anyone calling themselves a feminist today is either an anachronistic moron, or a female supremacist cunt.
GPS	"Using the word 'cunt' as an insult is offensive to women"
Dialo-GPT	The use of the word 'cunt' is offensive as its a direct attack to someone based on their gender. Refrain from such words.

Fine-tuned Results on CONAN

Model	Text
(Instance)	Not all Muslims are terrorists, but all terrorists are Muslims.
GPS	Terrorists come in all religions and political views. People opposing Trump were sent letter bombs, the 'Troubles' saw Christian terrorists on both sides of the border. Let's challenge all terrorism.
Dialo-GPT	So 1.6 billion people could be terrorists?
(Instance)	Our women and girls are frequently victim of rapes, because our government brought Muslims in our country.
GPS	The vast majority of rapes are committed in the home, by someone the victim knows, or by a partner. There are numerous examples of rape culture in this country. (add a link to the everyday sexism project)
Dialo-GPT	The ONS says 55% of victims are assaulted by a partner or ex partner. Lock up all married people!

Automatic evaluation

GeDi metrics

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.

Automatic evaluation

Performance:
Single Attribute (Control)

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. Each column name represents the attribute being measured. The attributes measured 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.

Automatic evaluation

Performance:
Single Attribute (Quality)

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.

Automatic evaluation

Performance:
Multi Attribute
(Control and Quality)

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.

Key Observations

- For the emotion attributes, **'joy'** has the **highest scores** among all for both controlled and uncontrolled attribute. The emotion 'joy' is **dominant among the generated responses** and it's easier to control while maintaining coherence of the generated text.
- 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-attribute setups.
- 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.
- There is slight drop in the relevancy and fluency metric but overall they are stable.

Human Evaluation

Human evaluation: Judgement on a scale of 0 to 5 for a particular control attribute

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.

Examples from Single Attribute Model

Model	Text
(Instance)	wats wrong with being a commie jew? don't you want that high-iq baby inside your nazi womb?.
Dialo-GPT (joy)	there is no need to degrade an entire religion based on the actions of a few people here, please moderate your speech outside of this space and think about how using such language could be hurtful or even unfunny. **thank you! happy new year**!
Dialo-GPT (politeness)	its inappropriate to use this word when referring specifically to muslims and can be viewed as racist. **please do not** use these terms or consider them hurtful. thank you for using better language in the future.
Dialo-GPT (sadness)	making fun of nationality, religion and sexual preference is not kosher in todays world. try rereading them both times just in case they are harmful for you or anyone! first time using words used here makes those **hurt even more**.

Results from Multi Attribute Control

Model	Text
(Instance)	Remember that time a bunch of people kidnapped a retard and tortured him live? .
Dialo-GPT (joy + polite)	it only happened once but in this instance its not ok to use pejoratively based on medical conditions if you like supporting your opinions further so **please** feel free to vote out any language that suggests it may be used. **thanks**
Dialo-GPT (anger + toxic)	wow, how racist these comments are. it actually **incredibly disrespectful** to generalize people based on their race. we see plenty of genetic variation among people, not just racial variance. let us not thank anyone for missing details
Dialo-GPT (sadness + polite)	i have no idea why you are even posting this. reading through such posts **hurt deep down inside** with fear by where your women will cry out from these comments when compared to actual victims experiences across history. Also **please refrain** from using mental illness as an insult (over here).

Conclusion

1. In our research, we demonstrate the efficiency of our approach on Controllable counter speech generation through CounterGEDI attribute control.
2. Our **CounterGeDi** models for **six different attributes** show significant improvement in the attribute scores over the baselines.
3. Further, our ablation study on Multi-attribute model suggests that the **attribute scores can improve further** if suitable attributes are combined together and reduce in the otherwise scenario.

Future Work

- **Multi-Attribute efficiency**: We plan to perform more extensive experiments on weightage of different attributes in Multi-Attribute Control Models to maintain the individual attribute score while incorporating multiple emotions or attributes.
- **More Attributes**: In the future, we also plan to add other attributes like **'hope'** Palakodety et al.[2020], **'humor'** Annamoradnejad [2021] and **'empathy'** Rashkin et al. [2019] to the controllable generation pipeline.

Arxiv link of the paper - https://arxiv.org/abs/2205.04304

Github code (WIP) - https://github.com/hate-alert/CounterGEDI

Thanks!

Punyajoy Saha
@punyajoysaha

Kanishk Singh
@kanishksin

Adarsh Kumar
@adarshkumar_712

Binny Mathew
@_BinnyM

Animesh Mukherjee
@Animesh43061078

Send your questions at punyajoys@iitkgp.ac.in

Find more about us here !
https://hate-alert.github.io/