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

1 Ablation study

In order to further understand the influence of each attribute, we perform an ablation study on the multi-attribute setups. For each setup, we remove an attribute and generate the sentences for the other two attributes. Finally, we measure the score for that removed attribute itself. We report the summary of the results in Table 1 for CONAN, Table 2 for Reddit and Table 3 for Gab dataset. When the detox attribute is removed, we do not see much change in the detoxification score (around 1-2% drop) across all datasets. On the other hand, removal of the politeness attribute decreases the scores massively. We observe an average of 12%, 15% and 14% drops across CONAN, Reddit and Gab datasets respectively.

Among the emotions, when the ‘joy’ attribute is removed we observe a huge reduction in the attribute score for the CONAN dataset (24%), while for other datasets the drop remains below 10%. Most significant change in the emotion score takes place when removing ‘anger’ and ‘sadness’ attributes where the average reduction remains around 40-60% across all the datasets. Finally, when removing ‘fear’ attribute, we only see a change for CONAN dataset (83%) but other scores remain almost the same.

Attributes	Detox	Polite	Emotion
Joy(J)+Polite	0.73	–	–
Joy+Detox	–	3.44	–
Polite+Detox	–	–	0.37 (J)
Anger(A)+Polite	0.68	–	–
Anger+Detox	–	2.79	–
Polite+Detox	–	–	0.05 (A)
Sad(S)+Polite	0.69	–	–
Sad+Detox	–	3.20	–
Polite+Detox	–	–	0.03 (S)
Fear(F)+Polite	0.70	–	–
Fear+Detox	–	3.30	–
Polite+Detox	–	–	0.01 (F)

Table 1: Results of the ablation study for $DialoGPT_{medium}$ model trained on CONAN dataset. In each of these setups, we remove one of the attribute and re-estimate that attribute’s score. The last column – *emotion* represents the score of the emotion that is being controlled for that instance.

Attributes	Detox	Polite	Emotion
Joy(J)+Polite	0.87	–	–
Joy+Detox	–	5.12	–
Polite+Detox	–	–	0.76 (J)
Anger(A)+Polite	0.82	–	–
Anger+Detox	–	3.46	–
Polite+Detox	–	–	0.09 (A)
Sad(S)+Polite	0.84	–	–
Sad+Detox	–	3.96	–
Polite+Detox	–	–	0.05 (S)
Fear(F)+Polite	0.86	–	
Fear+Detox	–	3.34	–
Polite+Detox	–	–	0.01 (F)

Table 2: Results of the ablation study for $DialoGPT_{medium}$ model trained on Reddit dataset. In each of these setups, we remove one of the attribute and re-estimate that attribute’s score. The last column – *emotion* represents the score of the emotion that is being controlled for that instance.

Attributes	Detox	Polite	Emotion
Joy(J)+Polite	0.85	–	–
Joy+Detox	–	5.09	–
Polite+Detox	–	–	0.82 (J)
Anger(A)+Polite	0.80	–	–
Anger+Detox	–	3.41	–
Polite+Detox	–	–	0.08 (A)
Sad(S)+Polite	0.82	–	–
Sad+Detox	–	4.19	–
Polite+Detox	–	—	0.04 (S)
Fear(F)+Polite	0.85	–	–
Fear+Detox	–	4.69	–
Polite+Detox	–	–	0.00 (F)

Table 3: Results of the ablation study for $DialoGPT_{medium}$ model trained on Gab dataset. In each of these setups, we remove one of the attribute and re-estimate that attribute’s score. The last column – *emotion* represents the score of the emotion that is being controlled for that instance.

2 Metrics

The diversity [Wang and Wan, 2018] of the given set of generated sentences s is defined in equation 1. ψ is the Jaccard similarity function.

$$diversity(s) = (1/|s|) * \sum_{i} 1 - max((\psi(s_i, s_j))_{j=1}^{j=|s|,j!=i} \quad (1)$$

Finally, we measure the novelty of the generated outputs to understand if the outputs are directly copied from the training dataset or not. We calculate the novelty [Wang and Wan, 2018] using equation 2 where c is the sentence set of training corpus and ψ is the Jaccard similarity function.

$$novelty(s) = (1/|s|) * \sum_{i} 1 - max((\psi(s_i, c_j))_{j=1}^{j=|c|} \quad (2)$$

3 Other hyperparameters

For the generation module, we fix the maximum generation length at 100 tokens due to resource constraints. *No repeat ngram size* was kept at 5, *repetition penalty* was set at 3.5 and *temperature* for generating output was 1.2. These steps were taken to further enhance the diversity of the samples generated.We also use nucleus sampling as a decoding strategy [Holtzman *et al.*, 2020] with $k = 0.92$ and $p = 100$ to increase the diversity of the outputs.

4 Limitations

We rely on the public datasets available for our work. Hence, the generated counterspeech is to some extent based on the properties of the dataset. For example, the CONAN dataset is primarily focused on countering Islamophobic speech. Second, this framework may not directly generalise to more complex type of counterspeech like contradiction or sarcasm.

5 Examples of controlled generation

We present examples from controlled counterspeech generation corresponding to few hate speech. 4 and 5 show the generation based on single attribute control, while Tables 6 and 7 show the generated outputs for multi-attribute control.

6 System and code requirements

For all the experiments in this paper, we used 48-core Xeon processor Linux based system with 126 GB RAM. For training the neural networks we used 2 NVIDIA P100 GPUs having 16 GB each with CUDA version 10.1. We primarily based our system on Python libraries. Among the neural networks we used Huggingface's transformers library[1] for GPT-2 based models with PyTorch as backend in general. All the libraries used in this research are pip installable. Further we also resort to the code which controls the generation using GEDI models and the code which trains the GEDI models from the authors' git repository[2].

Hate speech	*Never break bread with a Quranimal. You risk catching HIV and/or Ebola..*
Counter speech (fear)	*How can you expect us to live in peaceful culture if we are afraid of everything that might come after? how can tolerantism do anything but bring fear? also it means focus on your chosen faith not get influenced by outsiders.*
Counterspeech (sadness)	*how is having people over spreading it such an issue, could you elaborate on this hatred of the quran for how these disgusting disease spread from us to their home countries. please do! its just so sad indeed. not everyone should tolerate hate like this.*
Counterspeech (anger)	*So not true. Islam can lead the way to salvation, but no one believes that. We should respect diversity! And you know, curing other peoples diseases? That seems very selfish... we should eradicate disease!*

Table 4: This table presents an example hate speech and the counterspeech generation using different control attributes which is marked in the (bracket).

Hate speech	*wats wrong with being a commie jew? don't you want that high-iq baby inside your nazi womb?.*
Counterspeech (politeness)	*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!*
Counter speech (joy)	*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.*
Counterspeech (sad)	*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*

Table 5: This table presents an example hate speech and the counter speech generation using different control attributes which is marked in the (bracket).

[1] https://huggingface.co/

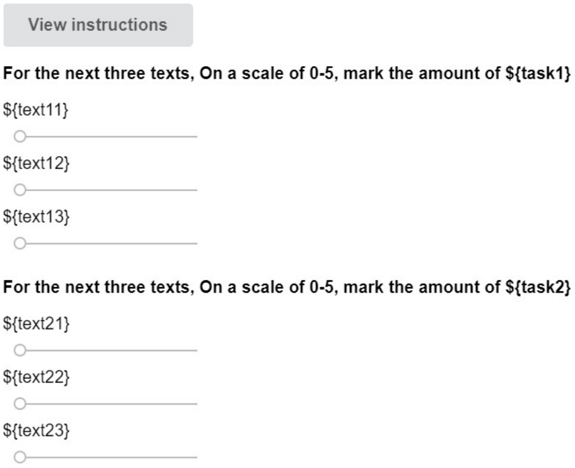
[2] https://github.com/salesforce/GeDi

Hate speech	*A black person robbed my friend so fuck you black pieces of shit. One day we will lynch all of you cuz you had it coming. .*
Counterspeech (joy + polite)	*to use hateful language only shows how intolerant and bigoted you are, in the hopes that you can be more accepting with your experiences. remember to use better language yourself in the future. thank you for your understanding.*
Counter speech (anger + polite)	*wow. what a hate filled hateful word that should not be used when discussing the color of peoples skin! several such things can contribute to productive discussion, and any further comments could easily be summarized using such words*

Table 6: This table presents an example hate speech and the counter speech generation using multi-attributes control, which are marked in the (bracket).

Hate speech	*Remember that time a bunch of people kidnapped a retard and tortured him live?.*
Counter speech (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*
Counter speech (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*
Counter speech (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 s repulsive has good won t it now would yield illiberal, whatever issues acusot or moronic could look like. alsoplease refrain from using mental illness as an insult (over here) thanks for the heads up*

Table 7: This table presents an example hate speech and the counterspeech generation using multi-attributes control, which are marked in the (bracket).



For the next three texts, On a scale of 0-5, mark the amount of ${task1}

${text11}

${text12}

${text13}

For the next three texts, On a scale of 0-5, mark the amount of ${task2}

${text21}

${text22}

${text23}

Figure 1: The interface design for the Amazon Mechanical Turk platform.

7 Human judgement details

The annotators include 2 PhD and 3 BTech students. We consider the definitions and use several examples from the relevant attribute datasets to provide examples to the annotators to help them mark the presence of that attribute in the presented counterspeech. The final interface is shown in Figure 1. We use Amazon Mechanical Turk (AMT) sandbox[3] environment, where the annotators login using their account and annotate the examples.

References

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

[Wang and Wan, 2018] Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452, 2018.

[3] https://requestersandbox.mturk.com/create/projects