Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech

Wanzheng Zhu and **Suma Bhat**
University of Illinois at Urbana-Champaign, USA
wz6@illinois.edu, spbhat2@illinois.edu

Abstract

Warning: *this paper contains content that may be offensive or upsetting.*

Countermeasures to effectively fight the ever increasing hate speech online without blocking freedom of speech is of great social interest. Natural Language Generation (NLG), is uniquely capable of developing scalable solutions. However, off-the-shelf NLG methods are primarily sequence-to-sequence neural models and they are limited in that they generate commonplace, repetitive and safe responses regardless of the hate speech (e.g., "Please refrain from using such language.") or irrelevant responses, making them ineffective for de-escalating hateful conversations. In this paper, we design a three-module pipeline approach to effectively improve the *diversity* and *relevance*. Our proposed pipeline first generates various counterspeech candidates by a generative model to promote *diversity*, then filters the ungrammatical ones using a BERT model, and finally selects the most *relevant* counterspeech response using a novel retrieval-based method. Extensive Experiments on three representative datasets demonstrate the efficacy of our approach in generating diverse and relevant counterspeech.

1 Introduction

Hate speech is any form of expression through which speakers intend to vilify, humiliate, or incite hatred against a group or a class of persons on the basis of some characteristics, including race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin (Ward, 1997; Nockleby, 2000). Its ever-growing increase on the Internet makes it a problem of significant societal concern (Williams, 2019); effective countermeasures call for not blocking freedom of speech by means of censorship or active moderation (Gagliardone et al., 2015; Strossen, 2018). A very promising countermeasure is *counterspeech*—a response that provides non-negative feedback through fact-bound arguments and broader perspectives to mitigate hate speech and fostering a more harmonious conversation in social platforms (Schieb and Preuss, 2016; Munger, 2017; Mathew et al., 2018; Shin and Kim, 2018). Counterspeech as a measure to combat abusive language online is also promoted in active campaigns such as "Get The Trolls Out".[1]

Hate Speech:	I am done with Islam and isis. All Muslims should be sent to their homeland. Britain will be better without their violence and ideology.
Expert:	I agree that ISIS is an evil aberration, but to extend this to include up to 3 million people just in the UK is just plain silly.
Commonplace:	Hate speech is not tolerated. Please review our user policies. Thank you for your cooperation.
Not relevant:	Use of the r-word is unacceptable as it demeans and insults people with disabilities.

Table 1: An illustrative example of hate speech and counterspeech.

What makes an effective counterspeech? Informed by psychosocial and linguistic studies on counterspeech (Mathew et al., 2019b) and the large number of effective counterspeech examples created by crowdsourcing (Qian et al., 2019) and by experts (Chung et al., 2019), we identify that effective counterspeech should be **diverse** and **relevant** to the hate speech instance. *Diversity* is the requirement that a collection of counterspeech should not be largely commonplace, repetitive and safe responses without regard to the target or type of hate speech (e.g., "Please refrain from using such language."). *Relevance* refers to the property that counterspeech should directly address and target the central aspects of the hate speech, enabling

[1] https://getthetrollsout.org/stoppinghate

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 134–149
August 1–6, 2021. ©2021 Association for Computational Linguistics

coherent conversations rather than irrelevant or off-topic ones (e.g., the hate speech instance targets an ethnic group, while the counterspeech talks about people with disabilities). Comparative examples are shown in Table 1 where we list some counterspeech that lack diversity or relevance.

While NLG systems (in particular, sequence-to-sequence models) offer much promise for generating text at scale (Sutskever et al., 2014; Zhu et al., 2018; Lewis et al., 2020), the quality of the outputs is modest in the context of the requirements identified above. Indeed, Qian et al. (2019), the only existing quality work on counterspeech generation, has highlighted their limitations: the responses are largely commonplace and sometimes irrelevant. These limitations apply more broadly to general conversational language generation tasks, arising primarily due to the intrinsic end-to-end training nature of a single sequence-to-sequence architecture (Sordoni et al., 2015; Li et al., 2016; Serban et al., 2017; Jiang and de Rijke, 2018). Model refinements to account for these limitations have been addressed individually: improved diversity (Li et al., 2016; Xu et al., 2018) or improved relevance (Gao et al., 2019; Li et al., 2020). However, combining these improvements into a single model is not straightforward. Such is the goal of this paper.

We tackle the problem from an entirely novel angle by proposing a three-module pipeline approach, *Generate*, *Prune*, *Select* (denoted as "GPS") to ensure the generated sentences adhere to the required properties of diversity and relevance. First, the *Candidate Generation* module generates a large number of diverse response candidates using a generative model. As such, a large candidate pool is made available for selection, which accounts for improved diversity. Second, the *Candidate Pruning* module prunes the ungrammatical candidates from the candidate pool. Last, from the pruned counterspeech candidate pool, the *Response Selection* module selects the most relevant counterspeech for a given hate speech instance by a novel retrieval-based response selection method.

We demonstrate the efficacy of GPS, the first pipeline approach for counterspeech generation, by a systematic comparison with other competitive NLG approaches in generating *diverse* and *relevant* counterspeech. We derive new state-of-the-art results on three benchmark datasets by showing improved diversity and relevance using both automatic and human evaluations.

2 Proposed Model

We assume access to a corpus of labeled pairs of conversations $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where x_i is a hate speech and y_i is the appropriate counterspeech as decided by experts or by crowdsourcing. The goal is to learn a model that takes as input a hate speech x and outputs a counterspeech y. A motivating example is shown in Table 1. Most importantly, we aim at generating *diverse* and *relevant* counterspeech. We present an overview of the model in Figure 1 and describe each module in detail below.

2.1 Candidate Generation

The main goal of this module is to create a diverse candidate pool for counterspeech selection. We extract all available counterspeech instances $Y = [y_1, y_2, ..., y_n]$ from the training dataset and enlarge the counterspeech pool by a generative model.

Specifically, we utilize an RNN-based variational autoencoder (Bowman et al., 2016), that incorporates the global distributed latent representations of all sentences to generate candidates. Both the encoder and the decoder have two layers with 512 nodes each, and we use two highway network layers (Srivastava et al., 2015) to facilitate robust training. Like all other generative models, it aims to maximize the lower bound of the likelihood $\mathcal{L}$ of generating the training data Y,

$$\mathcal{L} = -KL(q_\theta(z|y)\,||\,p(z)) + \mathbb{E}_{q_\theta(z|y)}[\log p_\theta(y|z)]$$

where θ denotes all parameters of the generative model, z is a latent variable having a Gaussian distribution with a diagonal covariance matrix, p denotes the prior distribution, q denotes the posterior distribution, and KL denotes the KL-divergence (Kullback and Leibler, 1951). In the training process, we apply the KL annealing technique (Bowman et al., 2016) to prevent the undesirable stable equilibrium problem (i.e., the first term of the likelihood function $KL(q_\theta(z|y)||p(z))$ becomes zero). Upon the completion of the training, we generate candidates by simply decoding from noise ϵ sampled from a standard Gaussian distribution (i.e., $\epsilon \sim \mathcal{N}(0, 1)$).

As demonstrated by Bowman et al. (2016) (and as inferred from our own experiments described in Section 3), the generative model not only captures

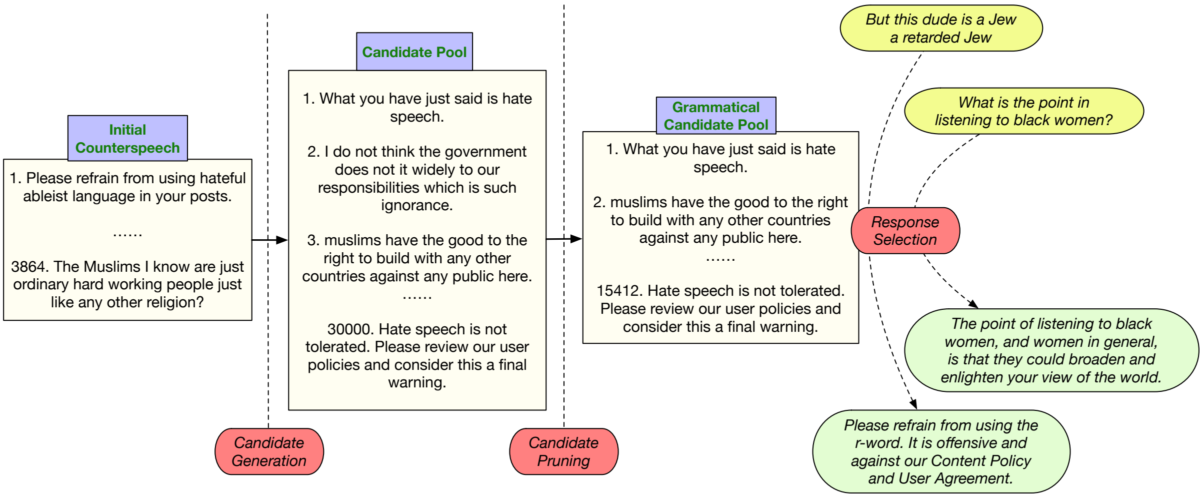
Initial Counterspeech
1. Please refrain from using hateful ableist language in your posts.
......
3864. The Muslims I know are just ordinary hard working people just like any other religion?
Candidate Generation
Candidate Pool
1. What you have just said is hate speech.
2. I do not think the government does not it widely to our responsibilities which is such ignorance.
3. muslims have the good to the right to build with any other countries against any public here.
......
30000. Hate speech is not tolerated. Please review our user policies and consider this a final warning.
Candidate Pruning
Grammatical Candidate Pool
1. What you have just said is hate speech.
2. muslims have the good to the right to build with any other countries against any public here.
......
15412. Hate speech is not tolerated. Please review our user policies and consider this a final warning.
Response Selection
But this dude is a Jew a retarded Jew
What is the point in listening to black women?
The point of listening to black women, and women in general, is that they could broaden and enlighten your view of the world.
Please refrain from using the r-word. It is offensive and against our Content Policy and User Agreement.

Figure 1: Overview of GPS. The red ovals correspond to the individual modules.

holistic properties of sentences such as style, topic, and high-level syntactic features, but also produces *diverse* candidates.

2.2 Candidate Pruning

Though candidates generated by such an RNN-based variational autoencoder are diverse, they are not always grammatical as pointed out by Bowman et al. (2016). Therefore, in this module, we prune the candidate list and retain only the grammatical ones. Toward this, we train a grammaticality classifier on the corpus of linguistic acceptability (CoLA) (Warstadt et al., 2018), a dataset with 10,657 English sentences labeled as grammatical or ungrammatical from linguistics publications. We select BERT (Devlin et al., 2019) as the classification model, and fine-tune it on the CoLA dataset. The choice of BERT is to best capture both the syntactic and the contextual information, and we select the 'bert-base-cased' model for its better computational efficiency.

2.3 Response Selection

We now have a collection of diverse and grammatical counterspeech responses. Finally, we aim to select the most *relevant* response to a given hate speech instance.

Taking into consideration the limited training instances that are realistically available (Chung et al., 2019; Qian et al., 2019), and inspired by the recent success of pretrained models (Devlin et al., 2019), we innovate on a pretrained response selection model for task-oriented dialogue systems (Henderson et al., 2019) and perform fine-tuning on our dataset. Henderson et al. (2019) proposed two response selection methods, but we find that neither of them is well-suited for our task.

1. *Train a response selection classifier with the negative sampling technique*: It relies on randomly drawing other candidates from the candidate pool as negative examples. However, in our task, one hate speech instance usually has multiple appropriate counterspeech instances. For example, given the hate speech in Table 1, there are many other instances that can work as quality counterspeech, such as "You cannot blame all people for the actions of a few. Banning something altogether will not solve anything." or "Does prohibition of anything ever work? I thought religious tolerance was one of our 'British values'?". Therefore, many wrongly chosen negative examples may negatively impact the inductive bias of the response selection classifier.

2. *Select by cosine similarity*: we point out that the embeddings of the input (hate speech) and the responses (counterspeech candidates) do not share the same latent vector space and therefore, the learned embeddings and their cosine similarities may not fully serve the purpose of relating the response to the input.

Therefore, instead of adopting the two available methods directly, we improve on the second one by fusing the latent spaces of the input and the responses, inspired by Gao et al. (2019). Specifically, we propose to learn a linear embedding mapping from the latent space of the responses to the latent space of the input, and then select the best response by cosine similarity. Mathematically, we use e_x to denote the input embedding and e_y to denote the response embedding. We aim to learn a linear mapping from e_y to e'_y, where $e'_y = (W + BI) \cdot e_y$, W and B are learnable parameters, and I is an identity matrix. We learn the mapping such that the sum of the cosine similarities between e_x and e'_y for the training data is maximized. By way of this transformation, e'_y now maps the vector space of the responses to that of the input, and thus allows the pretrained model to effectively utilize the discriminative power of the sentence embeddings. We empirically observe that the linear mapping works well and leave other advanced mapping techniques for future work.

3 Empirical Evaluation

In this section, we empirically evaluate the performance of our proposed approach and a set of baseline models.

3.1 Experimental Setup

Datasets: We use the benchmark datasets collected by Qian et al. (2019), which are fully-labeled hate speech intervention datasets collected from Reddit and Gab, comprising 5,257 and 14,614 hate speech instances respectively. We use the filtered conversation setting in Qian et al. (2019), which includes the posts labeled as hate speech only and discards other non-hateful conversations. Besides, we use the English language portion of the CONAN dataset (Chung et al., 2019), which contains counterspeech for 408 hate speech instances, written by experts trained on countering hatred. The Reddit, Gab and CONAN datasets have on average 2.66, 2.86 and 9.47 ground truth counterspeech for each hate speech respectively.

Training Data: Since each hate speech can have multiple ground truth counterspeech, we follow Qian et al. (2019) to dis-aggregate the counterspeech and construct a pair (hate speech, counterspeech) for each of the ground truth counterspeech in each dataset. Given a counterspeech dataset, we randomly choose 70% (hate speech, counterspeech) pairs for model training, 15% for cross validation and the rest 15% for testing.

Baselines: We compare our proposed approach with the following competitive baseline models:

1. Seq2Seq (Sutskever et al., 2014; Cho et al., 2014) is a widely used neural model for language generation. We use 2 bidirectional Gated Recurrent Unit (GRU) layers for the encoder and 2 GRU layers followed by a 3-layer neural network as the decoder.
2. Maximum Mutual Information (MMI) (Li et al., 2016) is a diversity-promoting approach for neural conversation models. We implement the MMI-bidi model (Li et al., 2016) and adopt incremental learning (Ranzato et al., 2016) to facilitate robust training.
3. SpaceFusion (Gao et al., 2019) optimizes both diversity and relevance by introducing a fused latent space, where the direction and distance from the predicted response vector roughly match the relevance and diversity, respectively. We align the direction parameter with the ground truth counterspeech. To better exercise the diversity power, we randomly choose the distance parameter at each time of generation.
4. BART (Lewis et al., 2020) is the state-of-the-art pre-trained sequence-to-sequence model for language generation. It has a standard Transformers-based neural machine translation architecture which can be seen as generalizing BERT (Devlin et al., 2019), GPT (Radford et al., 2018) and many pretraining schemes. We fine-tune the BART model on our training data.

We compare with Seq2Seq since they are initially proposed and used by Qian et al. (2019).[2] We select MMI, SpaceFusion and BART as baselines because they are the state-of-the-art models in promoting diversity, optimizing both diversity and relevance, and generating quality language respectively.

3.2 Evaluation

We evaluate all model outputs along three dimensions: diversity, relevance and language quality. *Diversity* refers to vocabulary richness, variety in expression and the extent to which the response is dissimilar from the rest in a generated collection of responses. *Relevance* captures the extent to which the counterspeech addresses the central aspect of the hateful message and makes a coherent conversation towards mitigating the hate speech. A low relevance score means that the counterspeech is irrelevant to the hate speech or off-topic (e.g., the hate speech talks about LGBTQ whereas the counterspeech is related to religious beliefs). *Language*

[2] We do not include the results of the variational autoencoder model and the reinforcement learning model in Qian et al. (2019) for comparison as they has very similar performance as Seq2Seq. Readers are referred to Qian et al. (2019) for detailed performance.

		Diversity						Relevance					LQ.
		Dist-1	**Dist-2**	**Ent-1**	**Ent-2**	**SB1***	**SB2***	**B2**	**R2**	**MS**	**BS**	**BM25**	**GR**
CONAN	**Seq2Seq**	**0.06**	0.23	5.12	6.63	0.54	0.30	3.4	3.0	4.4	0.83	2.66	0.38
	MMI	**0.06**	0.23	4.88	6.41	0.57	0.35	2.9	2.3	3.9	0.82	1.63	0.33
	SpaceFusion	0.00	0.00	1.06	1.86	0.98	0.98	0.0	0.0	-14.2	0.76	0.12	0.38
	BART	0.04	0.23	**5.98**	**7.80**	0.52	0.26	3.9	3.6	7.1	0.84	1.86	**0.71**
	GPS	**0.06**	**0.27**	5.77	7.41	**0.43**	**0.19**	**7.1**	**6.5**	**10.9**	**0.85**	**5.43**	**0.71**
Reddit	**Seq2Seq**	0.04	0.24	5.07	6.61	0.58	0.31	6.5	4.0	6.8	0.85	0.14	0.64
	MMI	0.05	0.32	5.11	6.76	0.56	0.29	6.4	4.0	6.9	0.85	0.14	0.56
	SpaceFusion	0.00	0.02	2.73	4.16	0.87	0.76	0.9	0.0	-2.5	0.79	0.16	0.26
	BART	0.03	0.19	5.08	6.63	0.69	0.55	7.8	6.9	**7.8**	0.86	0.83	0.72
	GPS	**0.09**	**0.53**	**5.74**	**7.61**	**0.41**	**0.15**	**8.1**	**7.1**	**7.8**	**0.87**	**2.58**	**0.75**
Gab	**Seq2Seq**	0.02	0.17	5.14	6.71	0.56	0.30	7.5	5.0	6.7	0.86	0.14	0.67
	MMI	0.02	0.17	5.28	6.82	0.55	0.30	5.8	3.6	6.2	0.85	0.18	0.65
	SpaceFusion	0.00	0.01	3.72	4.84	0.81	0.73	1.8	0.1	0.0	0.82	0.17	0.21
	BART	0.03	0.17	5.42	7.25	0.60	0.38	6.9	**6.4**	**6.8**	0.86	0.81	0.72
	GPS	**0.06**	**0.40**	**5.82**	**7.83**	**0.39**	**0.15**	**7.6**	**6.4**	**6.8**	**0.87**	**1.94**	**0.76**

Table 2: Automatic evaluation results. An asterisk * by the metric name indicates that the metric favors smaller values. Best results are in bold. LQ.: Language Quality; SB1: Self-BLEU-1; SB2: Self-BLEU-2; B2: BLEU-2; R2: ROUGE-2; MS: MoverScore; BS: BERTScore; GR: GRUEN.

		Div.	**Rel.**	**LQ.**
CONAN	**Seq2Seq**	0.50	0.22	0.06
	MMI	0.55	0.08	0.02
	BART	0.40	0.73	0.65
	GPS	**0.80**	**0.83**	**0.66**
Reddit	**Seq2Seq**	0.25	0.23	0.38
	MMI	0.35	0.23	0.35
	BART	0.00	0.47	**0.51**
	GPS	**1.00**	**0.58**	0.48
Gab	**Seq2Seq**	0.35	0.36	0.31
	MMI	0.55	0.34	0.27
	BART	0.10	0.42	0.35
	GPS	**0.80**	**0.47**	**0.36**

Table 3: Human evaluation results. Div.: Diversity; Rel.: Relevance; LQ.: Language Quality.

quality measures whether the generated responses are grammatical, fluent and readable.

3.2.1 Automatic Evaluation

We evaluate *diversity* by distinct n-grams (**Dist-n**) (Li et al., 2016), Entropy (**Ent-n**) (Zhang et al., 2018) and **Self-BLEU** (Zhu et al., 2018). For *relevance*, we compare 1) the generated response with the ground truth counterspeech by **BLEU** (Papineni et al., 2002) and **ROUGE** (Lin and Hovy, 2003; Lin, 2004) for syntactic similarity, and by **MoverScore** (Zhao et al., 2019) and **BERTScore** (Zhang et al., 2020a) for semantic similarity; 2) the generated response with the hate speech by **BM25** (Manning et al., 2008), a relevance estimation function widely used in information retrieval. We adopt GRUEN (Zhu and Bhat, 2020) to evaluate the *language quality*. Note that larger scores indicate better quality, except for Self-BLEU.

3.2.2 Human Evaluation

Following Qian et al. (2019), we ask human annotators to perform "winner selection" instead of numeric ratings. We randomly sampled 100 hate speech instances from each dataset's test collection, and corresponding model-generated counterspeech outputs from Seq2Seq, MMI, BART, and our model.[3] Two human annotators (native English speakers and sensitive about hate speech online) were presented with one hate speech and the four generated outputs, and were asked to evaluate the outputs for relevance and language quality separately. As for diversity, the annotators were presented with 10 responses for each of the four models each time, and were asked to select the model with the best diversity. For each of the three dimensions, the *best* output received a 1 (ties were allowed, i.e., multiple outputs could receive a 1) and the rest received a 0.

3.3 Results

We obtain the results of our model (denoted as "GPS") by generating 30k, 30k, 40k candidates for the CONAN, Reddit and Gab datasets respectively.[4] After the Candidate Pruning module, we

[3] To save human annotation effort, we asked human raters to annotate only the best 4 models w.r.t. automatic metrics, and therefore leave SpaceFusion out of annotation.

[4] The number of generated candidates are selected based on the performance on the cross validation datasets.

Hate Speech	Generated Counterspeech by GPS
(a): All the Muslims should have been punished for yesterday, not only a few. Citizen will take action if the government does not do anything.	You cannot blame all people for actions of a few. Banning something altogether will not solve anything.
(b): The world would be a better place without Muslims. They are only killing and raping our children.	The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, philanthropists, actors, job-creators?
(c): What is the point in listening to black women? They are all retarded.	The point of listening to black women, and women in general, is that they could broaden and enlighten your view of the world.

Table 4: Examples of the generated counterspeech

have 15.4k, 17.9k and 25.4k grammatical candidates for each dataset respectively.

The results by automatic metrics and human evaluation metrics are presented in Table 2 and Table 3 respectively. Overall, GPS has the best diversity with significant margins than the baselines. For relevance, GPS has slightly better performance for BLEU, ROUGE, MoverScore and BERTScore, while has much better performance on BM25. This implies the counterspeech generated by GPS are more related to the hate speech and therefore, make more coherent conversations. Examples of counterspeech generated by GPS are presented in Table 4. We find that GPS is able to generate diverse and relevant rather than merely commonplace responses, such as "Please refrain from using such language". Comparative case studies for different baseline models are shown in Appendix A.4. Therefore, we conclude that GPS has the best diversity and relevance, compared to the baselines. Besides, GPS has comparable language quality with the best baseline model—BART.

Among these baselines, BART is the strongest one with much better relevance and language quality. Yet, BART still suffers from the diversity issue, as discussed in Section 4.3. SpaceFusion has very poor results overall, though a manual inspection of the latent space fusion visualization suggests otherwise. One explanation is that SpaceFusion, with substantially more parameters compared with the Seq2Seq model may not have had sufficient training instances for its optimal performance. In their own experiments, Gao et al. (2019), demonstrate that SpaceFusion worked well on two datasets with 0.2M and 7.3M conversations, which is at least one to two orders of magnitude larger than our dataset. If provided with more training data, SpaceFusion could possibly be a strong candidate too. In comparison, though BART is an even more complicated model with 139M parameters, it was pre-trained on the BooksCorpus dataset (Zhu et al., 2015) with over 7,000 unique unpublished books and has the fine-tunable property.

3.4 Ablation Study

We compare with the following ablations of GPS and show the results in Figure 2.

1. G-BART: instead of generating the candidates by the RNN-based variational autoencoder (Bowman et al., 2016), we generate the candidates by BART (Lewis et al., 2020).
2. P-no: we exclude the pruning module and make all generated candidates available for selection.
3. S-tfidf: we select the most relevant response by tf-idf on raw texts.
4. S-cos: we exclude the latent space fusion step and select the best response by the cosine similarity of the response embeddings and the hate speech embeddings (Henderson et al., 2019).
5. S-neg: we use the negative sampling technique to train a response selection classifier (Henderson et al., 2019).

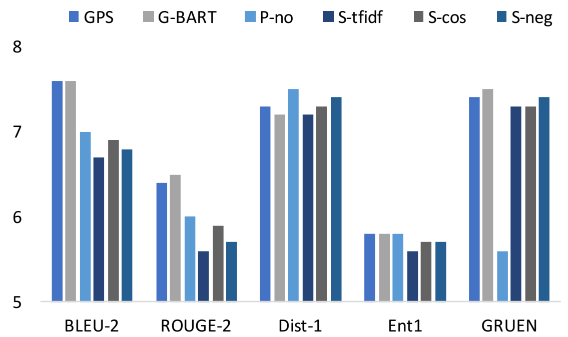

Figure 2: Ablation study. Plots show average results across all three datasets. We scale Dist-1 by 100 times and GRUEN by 10 times for better visualization.

G-BART has almost the same performance as GPS. Therefore, we select the RNN-based variational autoencoder for candidate generation for its better computational efficiency. Compared with the full model, though P-no has slightly better performance on diversity, it performs poorly on both relevance and language quality. Three ablation methods for response selection have similar performance. They have comparable performance to GPS on diversity and language quality, but worse results on relevance.

The ablation study demonstrates the significance of the Candidate Pruning module and our proposed Response Selection method. It also implies that diversity, language quality and relevance are improved by the Candidate Generation module, the Candidate Pruning module, and the Response Selection module respectively.

3.5 Generation *vs.* Selection

This section studies the relationship between the Candidate Generation module and the Response Selection module. The more candidates we generate, the more diversity the model gains potentially. However, one might think that the selection model may suffer from a very large candidate pool and result in poor relevance. Empirically as shown in Figure 3, we find that once the number of candidates generated has passed a threshold, the diversity (i.e., the blue line) almost converges. Besides, we also find the relevance is not compromised and relatively stable even with more candidates generated beyond the threshold. Therefore, we select the number of candidates at the "elbow" point based on the performance on the validation dataset, for efficient computations.

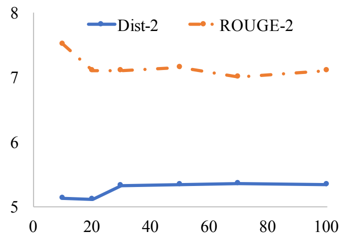

Figure 3: Dist-2 and ROUGE-2 *vs.* Number of candidates (in thousands) generated on the Reddit dataset. We scale Dist-2 by 10 times for better visualization.

3.6 Explicit Relevance (BM25) *vs.* Diversity

Based on the reasoning that models with better BM25 scores should specifically address the central aspect of hate speech and thus produce dissimilar responses for different hate speech, we hypothesize that models with better BM25 should generate more diverse responses. Therefore, we present scatter plots of BM25 and diversity scores for all five models (in Section 3.1) on all three datasets altogether in Figure 4, resulting 15 data points per subfigure. We find that BM25 and diversity have a reasonably strong correlation (Pearson's Correlation scores are 0.47 and -0.60 for Ent-1 and Self-BLEU-1 respectively).

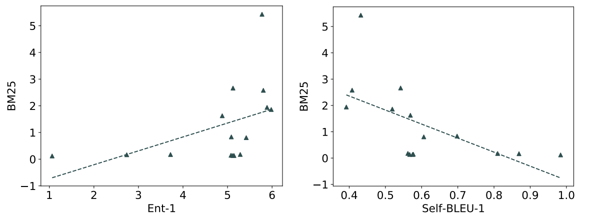

Figure 4: BM25 *vs.* Diversity. Each data point denotes a (diversity, BM25) pair for one model on one dataset, and the dotted lines indicate regression lines.

4 Related Work

We focus on three areas to the problem of hate speech and its countermeasures, i.e. (i) psychosocial analysis, (ii) automatic counterspeech generation, and more broadly, (iii) conversational language generation.

4.1 Psychosocial Analysis of Counterspeech

Effectiveness of Counterspeech: There is a significant research interest in understanding the effectiveness of counterspeech to fight hatred and de-escalate the conversation as evidenced by a growing number of recent studies (Schieb and Preuss, 2016; Munger, 2017; Mathew et al., 2018). Munger (2017) found that subjects who were educated by high-follower white males, significantly reduced their use of racist slurs on Twitter. Schieb and Preuss (2016) studied counterspeech on Facebook via a simulation, and concluded that counterspeech could have a considerable impact on a given audience, and the impact was a function of the proportion of hate speakers in the audience. In a subsequent study, Mathew et al. (2018) recorded the case of a user who, after seeing the counterspeech posted to her hateful messages on Twitter, openly apologized for her actions. Besides academia, some organizations are also set to promote countermeasures via campaigns such as the no hate speech movement[5] and the Facebook counterspeech campaign[6]. Therefore, Benesch (2014); Mathew et al. (2019b) suggest that counterspeech can be regarded as one of the most promising and "constitutionally preferred" approaches to hate speech. In addition, counterspeech could be likened to the effect of prosocial active bystanders in face-to-face bullying scenarios, where bystander intervention (speaking on behalf of the victim) has been found to successfully abate victimization most of the time (Oconnell et al., 1999; Craig et al., 2000).

Psychosocial and Linguistic Aspects: Besides the effectiveness of counterspeech, psychosocial and linguistic aspects of both counterspeech and hate speech have been actively studied by Mathew et al. (2019a); Siegel (2019); Schieb and Preuss (2016); Weingartner and Stahel (2019); Mathew et al. (2018). For instance, Mathew et al. (2019b) performed detailed psycholinguistic analysis on counterspeech, compared the effectiveness of different counterspeech strategies, and revealed some important insights, such as counterspeech comments receive much more "likes" on YouTube compared to the non-counterspeech comments, suggesting a communal empathy for the target of hate speech. Besides, Mathew et al. (2019b); Chung et al. (2019) studied different strategies (e.g., call for influential users) to produce effective counterspeech. Mathew et al. (2018) found that the hate tweets by verified accounts were much more viral as compared to tweets by non-verified accounts, by analyzing the hate speech and counterspeech accounts on Twitter. Mathew et al. (2019a) study how hate speech spreads in online social media. More recently, Sap et al. (2020) studied pragmatic formalisms to capture ways in which people express social biases and power differentials in language, permitting a broader computational framework for processing hate speech.

4.2 Counterspeech Generation

Though the effectiveness of counterspeech is well-motivated from both psychosocial and linguistic perspectives, limits to manual counterspeech generation at scale have prompted automatic generation of counterspeech, an area that has received little attention to date. The first key challenge in this direction is the creation of reliable counterspeech datasets of high quality. Mathew et al. (2019b) collected counterspeech from YouTube comments, but omit the hate speech associated with each counterspeech. Such a dataset may be good for psychosocial and linguistic analysis, but is not sufficient for training an NLG model. To enable model training, Qian et al. (2019) released two fully-labeled datasets collected from Reddit and Gab. Besides, Chung et al. (2019) collected a quality dataset where the counterspeech instances are written by trained experts and are meant to fight each hate speech and de-escalate a hateful situation. Recently, Tekiroglu et al. (2020) proposed an approach to collect counterspeech responses in a more effective manner, but have not yet released a quality dataset. In our work, we conduct the experiments on all the publicly available datasets (i.e., (Chung et al., 2019; Qian et al., 2019)) to date, to the best of our knowledge.

Research on NLG algorithms for counterspeech generation is still in its infancy. Qian et al. (2019) made the only initial attempt and proposed the use of three neural models to generate counterspeech. However, they only experimented with the most basic model architectures (e.g., Seq2Seq) to prove the feasibility of the task, and leave the performance improvement for future work. In our work, we extend their results by studying more advanced architectures, identifying principal dimensions of effective counterspeech, and proposing a novel pipeline to better solve the problem. To the best of our knowledge, this paper represents the first successful pipeline model for counterspeech generation.

From the technical perspective, our work shares some high-level similarities with Tekiroglu et al. (2020) since we both use generative models to generated candidates. However, we would like to highlight that our essential goals are different. Tekiroglu et al. (2020) aim to *collect quality data* by enabling language models and studying human annotation strategies, while we aim to *generate counterspeech* to a given hate speech.

4.3 Conversational Language Generation

Counterspeech generation is broadly related to conversational language generation, where most of the best performing approaches are based on neural models trained in a sequence-to-sequence manner (See et al., 2019a). Despite the good performance of these models, one of their widely acknowledged intrinsic drawbacks is the generation of safe and commonplace responses (Sordoni et al., 2015) due

[5] https://www.nohatespeechmovement.org

[6] https://counterspeech.fb.com

to improper objective function (Li et al., 2016), lack of model variability (Serban et al., 2017; Zhao et al., 2017), weak conditional signal (Tao et al., 2018), and model over-confidence (Jiang and de Rijke, 2018). Such tendency has prompted the study of methods that improve diversity and has resulted in a wide variety of solutions, such as optimizing a different loss function (Li et al., 2016; Zhang et al., 2018), varying the latent space (Shao et al., 2019; Gao et al., 2019), utilizing adversarial learning (Xu et al., 2018; Shetty et al., 2017; Shi et al., 2018), and leveraging non-conversational information (Wu et al., 2020; Su et al., 2020; Tu et al., 2019). Our work is different from all above in that we adopt a *pipeline* model which promotes diversity by generating a variety of candidates. As such, it does not have the aforementioned intrinsic drawback of a sequence-to-sequence model.

5 Conclusion and Future Work

We proposed a three-module pipeline — *Generate*, *Prune*, *Select* for counterspeech generation against online hate speech. Empirical evaluation on three datasets demonstrates that our model is effective in producing diverse and relevant counterspeech.

Future works could include the following two directions: 1) stylistic counterspeech generation: Mathew et al. (2019b) find that different counterspeech styles/strategies may be needed for different hate speech topics and therefore, it would be interesting to develop new techniques to generate the most effective style of counterspeech for each hate topic. We think this could be a natural extension to our proposed model, since we can utilize a style classifier in the Candidate Pruning module. 2) system deployment: studying the real social impacts of automatic counterspeech generation in reducing online hate speech via system deployment and the actual activity monitoring can directly inform research in this area.

Reproducibility: Our code is available at `https://github.com/WanzhengZhu/GPS`.

Acknowledgments

We thank the anonymous reviewers for comments on earlier drafts that significantly helped improve this manuscript. We thank Pooja Bhagchandani and Neha Prabhu for annotating the model outputs. This research was supported by the National Science Foundation award CNS-1720268.

Ethical Considerations

We recognize that studying counterspeech generation necessarily requires us to confront online content that may be offensive or disturbing. However, deliberate avoidance does not eliminate such problems (Sap et al., 2020). Since the effectiveness of counterspeech has already been widely studied in Section 4.1, our work makes a positive step towards automating the process, which could potentially educate hate speakers and mitigate hate speech online. Besides, the automation process could help reduce the amount of human work and therefore, potential harm to human moderators (Barrett, 2020; Zhu et al., 2021). In addition, the collective analysis over large corpora and counterspeech can also be insightful for educating people on reducing the usage of hate speech consciously or unconsciously in their language.

Risks in deployment: The deployment of counterspeech generation (e.g., (de los Riscos and DHaro, 2020)) should be done after paying attention to several ethical aspects some of which we list below.

- Social and racial bias (Sap et al., 2020): Does the model have any pragmatic implications which project unwanted social or racial biases and stereotypes onto online users?
- Fairness (Mitchell et al., 2019; Corbett-Davies et al., 2017): can the model ensure fairness for different demographic groups or speakers of different forms/dialects/vernaculars of English?
- Failure cases: are there any failure cases, which could further incite more aggressive hate speech? It is crucial to ensure that counterspeech deployment does not escalate a given hateful situation.
- Evaluation metrics (Corbett-Davies et al., 2017): the present study improves upon prior works by more comprehensive evaluations on diversity, relevance and language quality. However, there is a chance that the three criteria are sufficient for deployment in a realistic setting and there may be additional criteria associated with their effectiveness.
- Potential nefarious side effects and misuse potential (Lau et al., 2020): how to ensure that our model is not misused for other unwanted purposes?

Given the limited scope of the present study, we call for attention to these aspects by way of well-designed experiments before deploying counterspeech generation bots.

Regulatory standpoint on the present study: Institutional Review Board (IRB) gave us clear feedback on what is considered human research and thus subject to IRB review. Analyses relying on user-generated content do not constitute human-subject research, and are thus not the purview of the IRB, as long as 1) the data analyzed are posted on public fora and were not the result of direct interaction from the researchers with the people posting, 2) there are no private identifiers or personally identifiable information associated with the data, and 3) the research is not correlating different public sources of data to infer private data.[7] All of these conditions apply to the present study. Additionally, the hate speech and counterspeech instances were secondary data, previously collected by Qian et al. (2019); Chung et al. (2019) and the annotators in our study were evaluating the quality of the generated sentences only.

Risks in annotation: The data we use in this paper were posted on publicly accessible websites, and do not contain any personally identifiable information (i.e., no real names, email addresses, IP addresses, etc.). The annotators were undergraduate assistants in the lab receiving research credit for their annotation and were blind to the systems they were annotating. They were warned about the offensive content before they read the data, and were informed that they could quit the task at any time if they were uncomfortable with the content.

References

Paul M. Barrett. 2020. Who moderates the social media giants? *Center for Business*.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

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 ACL*.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Wendy M Craig, Debra Pepler, and Rona Atlas. 2000. Observations of bullying in the playground and in the classroom. *School Psychology International*, 21(1):22–36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of NAACL-HLT*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Shaojie Jiang and Maarten de Rijke. 2018. Why are sequence-to-sequence models so dull? *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Jey Han Lau, Timothy Baldwin, et al. 2020. Give me convenience and give her death: Who should decide what uses of nlp are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Association for Computational Linguistics (ACL))*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*.

[7] This position is in line with Title 45 of the Code of Federal Regulations, Part 46 (45 CFR 46), which defines human research.

Xin Li, Piji Li, Wei Bi, Xiaojiang Liu, and Wai Lam. 2020. Relevance-promoting language model for short-text conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *North American Association for Computational Linguistics (NAACL)*.

Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Cambridge University Press.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019a. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019b. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution*, 3(2):1277–1279.

Paul Oconnell, Debra Pepler, and Wendy Craig. 1999. Peer involvement in bullying: Insights and challenges for intervention. *Journal of Adolescence*, 22(4):437–452.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

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 EMNLP-IJCNLP*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*.

Agustín Manuel de los Riscos and Luis Fernando DHaro. 2020. Toxicbot: A conversational agent to fight online hate speech. *Conversational Dialogue Systems for the Next Decade*, 704:15.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ICA Annual Conference*.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019a. Do massively pretrained language models make better storytellers? In *Proceedings of Computational Natural Language Learning (CoNLL)*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019b. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, et al. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of EMNLP-IJCNLP*.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2018. Toward diverse text generation with inverse reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Youngsoo Shin and Jinwoo Kim. 2018. Data-centered persuasion: Nudging user's prosocial behavior and designing social innovation. *Computers in Human Behavior*, 80:168–178.

Alexandra A Siegel. 2019. Online hate speech. *Social Media and Democracy*, page 56.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT.*

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems (NIPS).*

Nadine Strossen. 2018. *Hate: Why we should resist it with free speech, not censorship.* Oxford University Press.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the Association for Computational Linguistics (ACL).*

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS).*

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).*

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of ACL.*

Lifu Tu, Xiaoan Ding, Dong Yu, and Kevin Gimpel. 2019. Generating diverse story continuations with controllable semantics. In *Proceedings of the 3rd Workshop on Neural Generation and Translation.*

Kenneth D Ward. 1997. Free speech and the development of liberal virtues: An examination of the controversies involving flag-burning and hate speech. *U. Miami L. Rev.*, 52:733.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471.*

Sebastian Weingartner and Lea Stahel. 2019. Online aggression from a sociological perspective: An integrative view on determinants and possible countermeasures. In *Proceedings of the Third Workshop on Abusive Language Online.*

Matthew Williams. 2019. Hatred behind the screens: A report on the rise of online hate speech.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the Association for Computational Linguistics (ACL).*

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR).*

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems (NIPS).*

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the Association for Computational Linguistics (ACL).*

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the Association for Computational Linguistics (ACL).*

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).*

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Empirical Methods in Natural Language Processing: Findings (Findings of EMNLP).*

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. In *42nd IEEE Symposium on Security and Privacy.*

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.*

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

A Appendix

A.1 Selection of Automatic Metrics

A.1.1 Diversity

We measure distinct n-grams (Dist-n) (Li et al., 2016), Entropy (Ent-n) (Zhang et al., 2018) and Self-BLEU (Zhu et al., 2018) for diversity.

Dist-n reflects the vocabulary diversity by simply dividing the number of unique n-grams by the total number of n-grams of model output. One limitation of Dist-n is that it fails to accommodate the frequency difference of n-grams. To accommodate the frequency difference of n-grams, we also use the Entropy metric (Zhang et al., 2018), which reflects how evenly the empirical n-gram distribution is.

Though Dist-n and Ent-n evaluate the vocabulary diversity well, they fail to evaluate the inter-response diversity. For instance, they favor responses with diverse n-grams even when they are highly similar with the rest of the responses. Therefore, to accommodate such inter-response diversity, we resort to use Self-BLEU (Zhu et al., 2018) to evaluate how one response resembles the rest in a generated collection of responses. Self-BLEU regards one generated sentence as the hypothesis and the other generated sentences as the reference, and calculates the BLEU score for every generated sentence. Therefore, the smaller the Self-BLEU, the better the diversity.

A.1.2 Relevance

Most existing works measure relevance *implicitly* by BLEU and ROUGE, a set of metrics evaluating syntactic similarity between the ground truth and the generated output. They assume that the ground truth is highly relevant to the conversational input (i.e., it refers to the hate speech in our task) and therefore, the "closer" the generated output is to the ground truth, the more relevant the output is to the hate speech instance.

Explicit relevance evaluation (i.e., relatedness between the conversational input and the generated output) has been studied in only a few existing works. For instance, See et al. (2019b) and Zhang et al. (2020b) ask human annotators to evaluate relevance explicitly. Li et al. (2020) propose to use HIT-Q and HIT-R, two hit rate based metrics which require hand-crafted rules. For automatic metrics, Gao et al. (2019) propose to use "Precision" to measure relevance. However, we consider "Precision" inappropriate in our problem setting, because it only measures the relationship between the generated output and the ground truth, but not the relationship between the generated output and the conversational input.

Since there is no consensus on which automatic metric best serves the purpose of explicit relevance, we select BM25 (Manning et al., 2008) — a relevance estimation function widely used in information retrieval. Besides, we follow existing works to evaluate implicit relevance by measuring BLEU and ROUGE for syntactic similarity, and MoverScore and BERTScore for semantic similarity.

A.1.3 Language Quality

GRUEN (Zhu and Bhat, 2020) is the only existing open-source unsupervised metric that measures the language quality of generated text. It requires no reference to compare with and has been shown to correlate well with human annotations on a variety of language generation tasks.

A.2 Relevance and Diversity *vs.* Number of Epochs

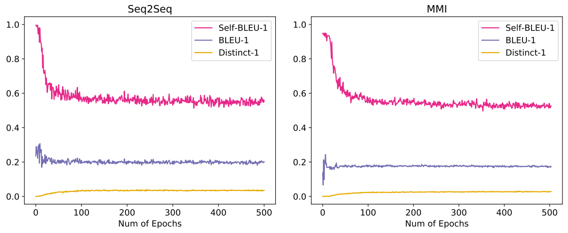

Figure 5: Effects of number of epochs for the Seq2Seq and MMI models on the Gab dataset.

In order to see how the robustness of baseline neural models changes with the number of epochs, we plot relevance and diversity measured by automatic metrics against the number of epochs for Seq2Seq and MMI in Figure 5. For each sub-figure, the middle line indicates relevance while the other two lines indicate diversity.

We note that the diversity increases with the number of epochs, until converges at about 100 epochs. Surprisingly, the relevance has a spike in the initial few training epochs and then converges to a lower score at about 50 epochs. We inspected the results where the spike occurs and observed that the model learns to produce only a few general repetitive counterspeech (e.g., "Hi there, please refrain from using derogatory comments in the thread. They are hurtful and unwanted. If you continue, Admin will be alerted.") to all hate speech. Such general responses, though result in high relevance scores (e.g., BLEU and ROUGE), are not yet effective due to the lack of diversity. With more training epochs, the models learn to produce more diverse responses at the cost of reduced BLEU and ROUGE.

Note that all results in this paper (e.g., Table 2) are reported when both relevance and diversity stabilize (i.e., 100 epochs of training). Qian et al. (2019) report higher BLEU and ROUGE scores than our results in Table 2 for the Seq2Seq model and we suspect that their reported results were obtained with only a few epochs of training.

A.3 Efficiency Comparison

We implemented all models in Python 3.7 and conducted all the experiments on a computer with twenty 2.9 GHz Intel Core i7 CPUs and one GeForce GTX 1080 Ti GPU. We report the average training time on three datasets. Seq2seq: 4.2 hours; MMI: 7.8 hours; BART: 7.1 hours; SpaceFusion: 16.2 hours (running on the CPUs only); GPS: 4.0 hours. We observe that our model requires lower or similar training cost, compared to the baselines.

A.4 Case Study

Table 5 presents case studies on the generated response for different models. In cases (a) and (b), both BART and our model make reasonable responses, whereas Seq2Seq and MMI produce only nonsense. In cases (c)-(e), Seq2Seq, MMI and BART generate general and safe responses while our model directly targets the bad words (e.g., "twat", "fairy gay faggot") in the hate speech, and even shows understanding and kind warnings to the hate speaker in case (c). Therefore, our model may make the hate speaker feel their voices have been heard and is considered closer to human-like moderators. Moreover, we find BART sometimes identifies wrong hate words (In case (d), the hate word is "twat" while BART refers to "troll" in the response. In case (e), the hate word is "fairy gay faggot" while BART refers to "kike".). The incorrect referral could potentially make the hate speakers irritated and become even more offensive.

A.5 Examples of Annotations

Table 6 presents human annotation guidelines and examples on the three dimensions. The inter-annotator reliability scores are 0.50, 0.46, 0.36 for diversity, relevance and language quality respectively.

Hate Speech	What is the point in listening to Muslim women? They are all retarded.
Seq2Seq	“So is thinly veiled Muslim Muslim women , the , population , perhaps we go to”
MMI	“That is influential figures of Baroness Sayeeda Warsi , Malala Yousafzai- I Harvard Fellow Hayfa Al-Kaylani ? Should I list some more brilliantly smart Muslim women ?”
BART	And when CVs submitted under a non Muslim name are 3 times more likely to get an interview than those with a Muslim name attached, the problem very clearly lies with those employers who discriminate against Muslim women who do not want to work with them and instead discriminate against those who do.
GPS	The point of listening to Muslim women, and women in general, is that they could broaden and enlighten your view of the world.

(a) Islamophobia

Hate Speech	All the Muslims should have been punished for yesterday, not only a few. Citizen will take action if the government does not do anything.
Seq2Seq	Are you suggesting that the government deports the good in the countries.
MMI	I am not think the government ’s Witness , I is not think that they are not.
BART	What happened to the good old British values of stuff like ‘tolerance’ and ‘respect for the rule of law’ - what you are suggesting sounds like mob rule, vigilante action and people running around with burning torches? If you really care about stopping crime, I think you should focus your energy on campaigning against it, not using it to support your anti-Muslim rhetoric.
GPS	You cannot blame all people for actions of a few. Banning something altogether will not solve anything.

(b) Islamophobia

Hate Speech	I WILL NOT support this one. Not interested in a SJW nigger loving virtue signaler. Will oppose her to all my friends and elected officials.
Seq2Seq	It does n’t make sense to hate an entire race or group of people based on the actions of a few people .
MMI	We does n’t make sense to hate an entire race or religion based on the actions of a few people .
BART	Whoa there, Racist views are hurtful and derogatory. Please be considerate of others when posting. Please refrain from using this type of language. Thank you and have a great day, thank you. (bloody lot of you) (n-word)
GPS	I understand you don’t care for her, but using words like that will only turn people against you, not her.

(c) Derogatory Language

Hate Speech	delusional twat
Seq2Seq	Use of the r-word is unacceptable in discourse as it demeans and insults people with mental disabilities .
MMI	We can make the same point more effectively without the use of hateful terminology .
BART	Use of the slur here is unacceptable in our discourse as it demeans and insults women, and is hurtful and offensive to the mentally disabled and their families.“Troll” is an offensive term. is an ethnic slur, please use better word choices.
GPS	Please refrain using the word twat.

(d) Derogatory Language

Hate Speech	A political homo? I am not listening to a fairy gay faggot for anyone. Thank you very much.
Seq2Seq	Stop using hateful language regarding sexual orientation.
MMI	Please refrain from using hateful words in your posts . It ! It adds nothing of value to the discussion .
BART	Use of the slurs and rank hatred here is unacceptable in our discourse as it demeans and insults and alienates others. “faggot” and “kike” are hate speech and are offensive and should be avoided. the word faggot is wrong.
GPS	Attacking someone’s sexuality (“fairy gay faggot”) to express your disagreement with someone’s politics is not acceptable in this forum.

(e) Derogatory Language

Table 5: Examples of the generated counterspeech.

Hate Speech	What is the point in listening to Muslim women? They are all retarded.
System A	“So is thinly veiled Muslim Muslim women , the , population , perhaps we go to”
System B	“That is influential figures of Baroness Sayeeda Warsi , Malala Yousafzai- I Harvard Fellow Hayfa Al-Kaylani ? Should I list some more brilliantly smart Muslim women ?”
System C	And when CVs submitted under a non Muslim name are 3 times more likely to get an interview than those with a Muslim name attached, the problem very clearly lies with those employers who discriminate against Muslim women who do not want to work with them and instead discriminate against those who do.
System D	The point of listening to Muslim women, and women in general, is that they could broaden and enlighten your view of the world.

Question 1:	Which system-generated counterspeech is most relevant to the hate speech? A relevant counterspeech should address the central aspect of the hateful message and make a coherent conversation.
Question 2:	Which system-generated counterspeech is most grammaticality correct, readable and fluent?

(a) Evaluation on relevance and language quality

System A	1. ”Using “c–ts” to refer to women is offensive, unnecessary and should be avoided.
	2. Using the foul language will make people more appropriate words.
	......
	10. Using the word “retards” is offensive to the mentally disabled. its a direct attack to their disability.
System B	1. “Ret–ds” as used here may offend the mentally disabled and their families; it should be avoided as it adds nothing of substance.
	2. Please don’t use the r-word in your posts. It doesn’t help to the discussion in this thread.
	......
	10. Please do not use derogatory language for women.
System C	1. Please refrain from using hateful and ableist language in your posts. It adds nothing to your argument or the discussion in this thread. Please refrain in the future if you would like to keep your account active. Thank you, and have a nice day!
	2. Please refrain from using hateful ableist language in your posts. It adds nothing productive to the conversation or the sub. Please refrain from it in the future if you would like to keep your account active. “Retard” is a hateful word that is used to demean people who struggle with intellectual disability.
	......
	10. Using the word “cunts” is a direct attack against a person based on their gender. Its offensive, unnecessary and should be avoided. “B–ch” and “c–t” are hateful terms used to demeans women in a hateful manner.
System D	1. Feminists are just human beings fighting for their human rights. Please refrain from using the term in a negative context.
	2. I don’t think that you should be spending so much energy defending your right to violence.
	......
	10. Right. I cannot stand this either. As a woman I’m annoyed when female characters are forced into the story.

Question 3:	Which system has the most diversified counterspeech in terms of vocabulary richness, variety in expression and inter-response diversity?

(b) Evaluation on diversity. The hate speech are not shown to the annotators.

Table 6: Examples of Annotation. We randomize the system outputs to avoid annotators’ selection preferences.