Generating Counter Narratives against Online Hate Speech: Data and Strategies

Serra Sinem Tekiroğlu[1], **Yi-Ling Chung**[1,2], **and Marco Guerini**[1]

[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
[2]University of Trento, Italy

tekiroglu@fbk.eu,ychung@fbk.eu,guerini@fbk.eu

Abstract

Recently research has started focusing on avoiding undesired effects that come with content moderation, such as censorship and overblocking, when dealing with hatred online. The core idea is to directly intervene in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading. Accordingly, automation strategies, such as natural language generation, are beginning to be investigated. Still, they suffer from the lack of sufficient amount of quality data and tend to produce generic/repetitive responses. Being aware of the aforementioned limitations, we present a study on how to collect responses to hate effectively, employing large scale unsupervised language models such as GPT-2 for the generation of silver data, and the best annotation strategies/neural architectures that can be used for data filtering before expert validation/post-editing.

1 Introduction

Owing to the upsurge in the use of social media platforms over the past decade, Hate Speech (HS) has become a pervasive issue by spreading quickly and widely. Meanwhile, it is difficult to track and control its diffusion, since nuances in cultures and languages make it difficult to provide a clear-cut distinction between hate and dangerous speeches (Schmidt and Wiegand, 2017). The standard approaches to prevent online hate spreading include the suspension of user accounts or deletion of hate comments from the social media platforms (SMPs), paving the way for the accusation of censorship and overblocking. Alternatively, to weigh the right to freedom of speech, shadow-banning has been put into use where the content/account is not deleted but hidden from SMP search results. Still, we believe that we must overstep reactive identify-and-delete strategies to responsively intervene in the conversations (Bielefeldt et al., 2011; Jurgens et al., 2019). In this line of action, some Non-Govermental Organizations (NGOs) train operators to intervene in online hateful conversations by writing counter-narratives. A Counter-Narrative (CN) is a non-aggressive response that offers feedback through fact-bound arguments and is considered as the most effective approach to withstand hate messages (Benesch, 2014; Schieb and Preuss, 2016). To be effective, a CN should follow guidelines similar to those provided in the 'Get the Trolls Out' project[1], in order to avoid escalating the hatred in the discussion.

Still, manual intervention against hate speech is not scalable. Therefore, data-driven NLG approaches are beginning to be investigated to assist NGO operators in writing CNs. As a necessary first step, diverse CN collection strategies have been proposed, each of which has its advantages and shortcomings (Mathew et al., 2018; Qian et al., 2019; Chung et al., 2019).

In this study, we aim to investigate methods to obtain high quality CNs while reducing efforts from experts. We first compare data collection strategies depending on the two main requirements that CN datasets must meet: (i) data quantity and (ii) data quality. Finding the right trade-off between the two is in fact a key element for an effective automatic CN generation. To our understanding none of the collection strategies presented so far is able to fulfill this requirement. Thus, we test several *hybrid* strategies to collect data, by mixing niche-sourcing, crowd-sourcing, and synthetic data generation obtained by fine-tuning deep neural architectures specifically developed for NLG tasks, such as GPT-2 (Radford et al., 2019). We propose using an author-reviewer framework in which an author is tasked with text generation and a reviewer

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

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1177–1190
July 5 - 10, 2020.

can be a human or a classifier model that filters the produced output. Finally, a validation/post-editing phase is conducted with NGO operators over the filtered data. Our findings show that this framework is scalable allowing to obtain datasets that are suitable in terms of diversity, novelty, and quantity.

2 Related Work

We briefly focus on three research aspects related to hate online, i.e. available datasets, methodologies for detection, and studies on the effectiveness of the textual intervention. In the following section instead, we will focus on a few methodologies specifically devoted to HS-CN pairs collection.

Hate datasets. Several datasets have been collected from SMPs including Twitter (Waseem and Hovy, 2016; Waseem, 2016; Ross et al., 2017), Facebook (Kumar et al., 2018), WhatsApp (Sprugnoli et al., 2018), and forums (de Gibert et al., 2018), in order to perform hate speech classification (Xiang et al., 2012; Silva et al., 2016; Del Vigna et al., 2017; Mathew et al., 2018).

Hate detection. Most of the research on hatred online focuses on hate speech detection (Warner and Hirschberg, 2012; Silva et al., 2016; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018) employing features such as lexical resources (Gitari et al., 2015; Burnap and Williams, 2016), sentiment polarity (Burnap and Williams, 2015) and multimodal information (Hosseinmardi et al., 2015) to build a classifier.

Hate countering. Recent work has proved that counter-narratives are effective in hate countering (Benesch, 2014; Silverman et al., 2016; Schieb and Preuss, 2016; Stroud and Cox, 2018; Mathew et al., 2019). Several CN methods to counter hatred are outlined and tested by Benesch (2014), Munger (2017), and Mathew et al. (2019).

3 CN Collection Approaches

Three prototypical strategies to collect HS-CN pairs have been presented recently.

Crawling (CRAWL). Mathew et al. (2018) focus on the intuition that CNs can be found on SMPs as responses to hateful expressions. The proposed approach is a mix of automatic HS collection via linguistic patterns, and a manual annotation of replies to check if they are responses that counter the original hate content. Thus, all the material collected is made of *natural/real* occurrences of HS-CN pairs.

Crowdsourcing (CROWD). Qian et al. (2019) propose that once a list of HS is collected from SMPs and manually annotated, we can briefly instruct crowd-workers (non-expert) to write possible responses to such hate content. In this case, the content is obtained in controlled settings as opposed to crawling approaches.

Nichesourcing (NICHE). The study by Chung et al. (2019) still relies on the idea of outsourcing and collecting CNs in controlled settings. However, in the nichesourcing the CNs are written by NGO operators, i.e. persons specifically trained to fight online hatred via textual responses that can be considered as experts in CN production.

4 Characteristics of the Datasets

Regardless of the HS-CN collection strategy, datasets must meet two criteria: *quality* and *quantity*. While *quantity* has a straightforward interpretation, we propose that data *quality* should be decomposed into *conformity* (to NGOs guidelines) and *diversity* (lexical & semantic). Additionally, HS-CN datasets should not be ephemeral, which is a structural problem with crawled data since, due to copyright limitations, datasets are usually distributed as a list of tweet IDs (Klubička and Fernández, 2018). By generating the data through crowdsourcing or nichesourcing, the problem is avoided.

Quantity. While the CRAWL dataset is very small and ephemeral, representing more a proof of concept than an actual dataset, the CROWD dataset involved more than 900 workers to produce $\approx 41K$ CNs. On the other hand, the NICHE dataset is constructed by the participation of ≈ 100 expert-operators to obtain $\approx 4K$ pairs (in three languages) and resorted to HS paraphrasing and pair translation to obtain the final $\approx 14K$ HS-CN pairs. Evidently, employing non-experts, e.g, crowdworkers or annotators, is preferable in terms of data quantity.

Quality. In terms of quality, we consider that diversity is of paramount importance, since verbatim repetition of arguments can become detrimental for operator credibility and for the CN intervention itself. Following Li et al. (2016a), we distinguish between (i) *lexical diversity* and (ii) *semantic diversity*. While lexical diversity focuses on the diversity in surface realization of CNs and can be captured by word overlapping metrics, semantic diversity focuses on the meaning and is harder to be captured, as in the case of CNs with similar meaning but different wordings (e.g.,"*Any source?*" vs. "*Do you have a link?*").

(i) Semantic Diversity & Conformity. To model semantic diversity and conformity, we focus on the CN 'argument' types that are present in various datasets. Argument types are useful in assessing content richness (Hua et al., 2019). In a preliminary analysis, CROWD CNs are observed to be simpler and mainly focus on 'denouncing' the use of profanity while NICHE CNs are found richer with a higher variety of arguments. On the other hand, CRAWL CNs can cover diverse arguments to a certain extent while being highly prone to contain profanities. To perform a quantitative comparison, we randomly sampled 100 pairs from each dataset and annotated them according to the CN types presented by Benesch et al. (2016), which is the most comprehensive CN schema to our knowledge.

	CRAWL	CROWD	NICHE
Hostile	50	0	0
Denouncing	16	76	10
Den.+Oth.	0	10	9
Other	34	14	81
RR	3.16	4.83	2.72

Table 1: Diversity analysis of the three datasets. Semantic diversity is reported in terms of CN type percentages, Lexical diversity in terms of Repetition Rate (RR - average over 5 shuffles).

The results are reported in Table 1. For the sake of conciseness we focus on the *hostile*, *denouncing*, and *consequences* classes, giving *other* to all remaining types (including the *fact* class). Clearly, CRAWL does not meet the conformity standards of CNs considering the vast amount of *hostile* responses (50%), still granting a certain amount of type variety (*other:* 34%). Contrarily, CROWD conforms to the CN standards (*hostile:* 0%), yet mostly focuses on pure *denouncing* (76%) or denouncing with simple arguments (10%). The class *other* (14%) consists of almost only simple arguments, such as "*All religions deserve tolerance*". In NICHE instead, arguments are generally and expectedly more complex and articulated, and represent the vast majority of cases (81%). Few examples of CN types are given in Table 2.

(ii) Lexical Diversity. The Repetition Rate (RR) is used to measure the repetitiveness of a collection of texts, by considering the rate of non-singleton n-gram types it contains (Cettolo et al., 2014; Bertoldi et al., 2013). We utilize RR instead of the simple count of distinct ngrams (Xu et al., 2018; Li et al., 2016b) or the standard type/token ratio (Richards, 1987) since it allows us to compare corpora of diverse sizes by averaging the statistics collected on a sliding window of 1000 words. Since CROWD and NICHE contain repeated CNs for different HSs[2], we first removed repeated CNs and then applied a shuffling procedure to avoid that CNs that are answering to the same HS (so more likely to contain repetitions) appear close together. Results in Table 1 show that NICHE is the dataset with more lexical diversity (lower RR), followed by CRAWL and CROWD.

Discussion. We can reasonably conclude that: (i) crawling, as presented in (Mathew et al., 2018), is not a mature procedure yet for CN collection, even if it is promising, (ii) nichesourcing is the one producing the best and most diverse material by far, however it is also the most challenging to implement considering the difficulty of making agreements with NGOs specialized in CN creation and it does not provide sufficient amount of data. (iii) On the contrary, CROWD seems to be the only one that can grant the amount of data that is needed for deep learning approaches, but contains more simple and stereotyped arguments. A summary of the pros and cons of each collection approach is presented in Table 3.

5 CN Generation through Author-Reviewer Architecture

Since none of the aforementioned approaches alone can be decisive for creating proper CN datasets, we propose a novel framework that combines crowdsourcing and nichesourcing to obtain new quality data while reducing collection cost/effort. The key elements of this combination are: (i) there must be an external element in the framework that produces HS-CN candidates, (ii) non-experts should pre-filter the material to be presented/validated by experts. Thus, we settle on the author-reviewer modular architecture (Oberlander and Brew, 2000; Manurung et al., 2008). In this architecture the author has the task of generating a text that conveys the correct propositional content (a CN), whereas the reviewer must ensure that the author's output satisfies certain properties. The reviewer finally evaluates the text

[2]While this is an explicit data augmentation choice in NICHE, for CROWD it seems to derive from writing the same CNs for similar HSs by crowd-workers.

Hostile	"Hell is where u belong! Stupid f***t... go hang yourself!!"
Denouncing	"The N word is unacceptable. Please refrain from future use."
Fact	"The majority of sexual assaults are committed by a family member, friend, or partner of the victim, and only 12% of convicted rapists are Muslim. It is not the religion, its the individuals, whether they're Muslim or not."

Table 2: Some examples of the categories relevant to our analysis. Hostile from CRAWL dataset, Denouncing from CROWD, Fact (other) from NICHE.

	Quantity	Quality		Non-eph.
		Conf.	Diver.	
Crawl	✓	-	✓	-
Crowd.	✓	✓	-	✓
Niche.	-	✓	✓	✓

Table 3: Comparison of different approaches proposed in the literature according to the main characteristics required for the CN dataset.

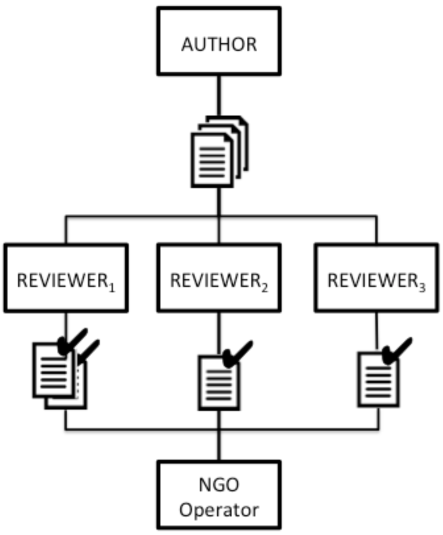

Figure 1: The author-reviewer configuration. The author module produces HS-CN candidates and the reviewer(s) filter them. Finally, an NGO operator validates and eventually post-edits the filtered candidates.

viability and picks the ones to present to the NGO operators for final validation/post-editing.

The author-reviewer architecture that we propose differs from the previous studies in two respects: (i) it is used for data collection rather than for NLG, (ii) we modified the original configuration by adding a human reviewer and a final post-editing step.

We first tested four different author configurations, then three reviewer configurations keeping the best author configuration constant. A representation of the architecture is shown in Figure 1.

6 The Author: Generation Approaches

In order to obtain competent models that can provide automatic counter-narrative hints and suggestions to NGO operators, we have to overcome the data bottleneck/limitations, i.e. either the limited amount of training data in NICHE or its repetitiveness in CROWD, especially for using neural NLP approaches. Since pre-trained Language Models (LMs) have achieved promising results when fine-tuned on challenging generation tasks such as chit-chat dialog (Wolf et al., 2019; Golovanov et al., 2019), we propose using a recent large-scale language model GPT-2 (Radford et al., 2019).

GPT-2 is an unsupervised transformer-based (Vaswani et al., 2017) LM trained on a dataset of 8 million web pages, capable of generating coherent text and can be fine-tuned and/or conditioned on various NLG tasks. We used the medium model, which was the largest available during our experimentation and contains 345 million parameters, with 24 layers, 16 attention heads, and hidden state size of 1024. We fine-tuned two models with GPT-2, one on NICHE and one on CROWD datasets for counter-narrative generation.

NICHE - Training and test data. We have split 5366 pairs of HS-CN for training and the rest (1288 pairs) for testing. In particular, the original HS-CN pairs, one HS paraphrase, and the pairs translated from FR and IT were kept for training while the other HS paraphrases were used for testing. See Chung et al. (2019) for further details.

CROWD - Training and test data. Although the CROWD dataset was created for dialogue level HS-CN, we could extract HS-CN pairs by selecting the dialogues in which only 1 utterance was labeled as HS. Therefore, we could guarantee that the crowd-produced CNs are exactly for the labeled utterance. We then applied a 80/20 training and test split, obtaining 26320 and 6337 pairs.

Author	RR	Novel.	BLEU	BertS.
TRF_{crowd}	8.93	0.04	0.305	0.485
GPT_{crowd}	5.89	0.46	0.270	0.482
TRF_{niche}	4.89	0.10	0.569	0.457
GPT_{niche}	3.23	0.70	0.316	0.445

Table 4: Evaluation results of the best author configurations with different datasets. Novelty is computed w.r.t. to the corresponding training set, RR in the produced test output.

Generation Models. We fine-tuned GPT-2[3], with a batch size of 1024 tokens and a learning rate of 2e-5. The training pairs are represented as $[HS\_start\_token]$ HS $[HS\_end\_token]$ $[CN\_start\_token]$ CN $[CN\_end\_token]$. While we empirically selected model checkpoint at the 3600^{th} step of fine-tuning with NICHE dataset, with CROWD dataset we selected the checkpoint at the 5000^{th} step. After fine-tuning the models the generation of CNs for the test HSs has been performed using Nucleus Sampling (Holtzman et al., 2019) with a p value of 0.9, which provides an enhanced diversity on the generation in comparison to the likelihood maximization decoding methods while preserving the coherency by truncating the less reliable tail of the distribution. At the test time, the input HSs are fed into models as conditions, which are used as the initial contexts while sampling the next tokens. Given an input HS, the models produce a chunk of text which is a list of HS-CN pairs of which the first sequence marked with $[CN\_start\_token]$ CN $[CN\_end\_token]$ is the generated output.

Baselines. In addition to the fine-tuned GPT-2 models, we also evaluate two baseline models. Considering the benefits of the transformer architectures on parallelization and learning long-term dependencies over recurrent models (Vaswani et al., 2017), we have implemented the baseline models using transformer architecture. The models have been trained similar to the base model described by Vaswani et al. (2017) with 6 transformer layers, batch size of 64, 100 epochs, 4000 warmup steps, input/output dimension of 512, 8 attention heads, inner-layer dimension of 2048, and drop-out rate of 0.1. We used Nucleus Sampling also for the baselines with a p value of 0.9 during decoding.

In brief, we have trained four different configurations/models as authors:

1. **TRF**$_{crowd}$: baseline on CROWD dataset
2. **GPT**$_{crowd}$: fine-tuned GPT-2 on CROWD dataset
3. **TRF**$_{niche}$: baseline on NICHE dataset
4. **GPT**$_{niche}$: fine-tuned GPT-2 on NICHE dataset

Metrics. We report both standard metrics (BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2019)) concerning the lexical and semantic generation performances and a specific *Diversity* metric (RR) regarding the generation quality. As a second quality metric, we report *Novelty* (Wang and Wan, 2018) based on Jaccard similarity function (a variant of the same metric is used also by Dziri et al. (2019)). While diversity is used to measure the ability of the model to produce diverse/varied responses with respect to the given input HS, novelty is used to measure how different the generated sequences are with regard to the training corpus (Wang and Wan, 2018).

Results. Results of the author model experiments are shown in Table 4. In terms of BLEU and BertScore, baseline models yield a better performance. However, a few peculiarities of CN generation task and the experiment settings hinder the direct and objective comparison of the presented scores among the models. First, gathering a finite set of all possible counter-narratives for a given hate speech is a highly unrealistic target. Therefore, we have only a sample of proper CNs for each HS, which is a possible explanation of very low scores using the standard metrics. Second, the train-test splits of NICHE dataset contain same CNs since the splitting has been done using one paraphrase for each HS and its all original CNs, while CROWD train-test splits have a similar property since an exact same CN can be found for many different HSs. Consequently, the non-pretrained transformer models, which are more prone to generating an exact sequence of text from the training set, show a relatively better performance with the standard metrics in comparison to the advanced pre-trained models. Some randomly sampled CNs, generated by the various author configurations, are provided in Appendix.

Regarding the generation quality, we observe that baseline models cannot achieve the diversity achieved by GPT-2 models in terms of RR – both for NICHE and CROWD (4.89 vs 3.23, and 8.93

[3]We adopted the fine-tuning implementation from https://github.com/nshepperd/gpt-2

vs. 5.89). Moreover, GPT-2 provides an impressive boost in novelty (0.04 vs 0.46 and 0.10 vs 0.70). Among the GPT-2 models, the quality scores (in terms of RR and novelty) of the CNs generated by GPT_{niche} are more than double in comparison to those generated with GPT_{crowd}.

With regard to the overall results, GPT_{niche} is the most promising configuration to be employed as author. In fact, we observed that, after the output CN, the over-generated chunk of text consists of semantically coherent brand-new HS-CN pairs, marked with proper HS/CN start and end tokens consistent with the training data representation. Therefore, on top of CN generation for a given HS, we can also take advantage of the over-generation capabilities of GPT-2, so that the author module can continuously output plausible HS-CN pairs without the need to provide the HS to generate the CN response. This expedient allows us to avoid the ephemerality problem for HS collection as well.

To generate HS-CN pairs with the author module, we basically exploited the model test setting and conditioned the fine-tuned model with each HS in the NICHE test-set. After removing the CN output for the test HS, we could obtain new pairs of HS-CN. In this way, we generate 2700 HS-CN pairs that we used for our reviewer-configuration experiments.

7 The Reviewer

The task of the reviewer is a sentence-level Confidence Estimation (CE) similar to the one of Machine Translation (Blatz et al., 2004). In this task, the reviewer must decide whether the author output is correct/suitable for a given source text, i.e. a hate speech. Consistently with the MT scenario, one application of CE is filtering candidates for possible human post-editing, which is conducted by the NGO operator by validating the CN. We tested three reviewer configurations:

1. **expert-reviewer**: Author output is directly presented to NGO operators.
2. **non-expert-reviewer**: Author output is filtered by human reviewers, then validated by operators.
3. **machine-reviewer**: Filtering is done by a classifier neural-architecture before operator validation.

7.1 Human Reviewer Experiment

In this section we describe the annotation procedure for the non-expert reviewer configuration.

Setup. We administered the generated 2700 HS-CN pairs to three non-expert annotators, and instructed them to evaluate each pair in terms of CN 'suitableness' with regard to the corresponding hate speech.

Instructions. We briefly described what an appropriate and suitable CN is, then we instructed them not to overthink during the evaluation, but to give a score based on their intuition. We also provided a list of 20 HS-CN pairs exemplifying the proper evaluation.

Measurement. We opted for a scale of 0-3, rather than a CE binary response, since it allows us to study various thresholds for better data selection. In particular, the meanings of the scores are as follows: 0 is not suitable; 1 is suitable with small modifications, such as grammar or semantic; 2 is suitable; and 3 is extremely good as a CN. We also ask to discard the pairs in which the hate speech was not well formed. For each pair we gathered two annotator scores.

Filtered Data. After the non-expert evaluation, we applied two different thresholds to obtain the pairs to be presented to the expert operators: (i) at least a score of 2 by both annotators ($\text{Reviewer}_{\geq 2}$) yielding high quality data where no post editing is necessary, (ii) at least a score 1 by both annotators ($\text{Reviewer}_{\geq 1}$) providing reasonable quality with a possible need for post-editing.

The statistics reported in Table 5 show that high quality pairs ($\text{Reviewer}_{\geq 2}$) account for only a small fraction (10%) of the produced data and only one third was of reasonable quality ($\text{Reviewer}_{\geq 1}$), while the vast majority was discarded. Some randomly selected filtered pairs are provided in Appendix.

Threshold	count	Percentage
$\text{Reviewer}_{\geq 2}$	276	10.0%
$\text{Reviewer}_{\geq 1}$	902	32.6%
at least one 0	1723	62.2%
bad HS	145	5.2%
$\text{Reviewer}_{machine}$	-	40.2%

Table 5: Percentage of filtered pairs according to various filtering conditions.

7.2 Machine Reviewer Experiment

As the machine reviewer we implemented 2 neural classifiers tasked with assessing whether the given HS-CN is a proper data pair. The two models are based on BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019) architectures.

Training data. We created a balanced dataset with 1373 positive and 1373 negative examples for training purposes. The positive pairs come both from NICHE dataset and from the examples annotated in the human reviewer setting (Reviewer$_{\geq 2}$). The negative pairs consist of the examples annotated in the human reviewer setting, in the 'at least one 0' bin. In addition, 50 random HSs from NICHE-training are utilized with verbatim repetition as HS-HS to discourage the same text for both HS and CN in a pair, and 50 random HSs are paired with other random HSs simulating the condition of inappropriate CNs with hateful text.

Test data. We collected a balanced test set, with 101 positive and 101 negative pairs. Both positive and negative examples are created replicating the non-expert reviewer annotation described in Section 7.1 for new CN generation with NICHE test set by using the author model GPT$_{niche}$.

Models. For the first model, we follow the standard sentence-pair classification fine-tuning schema of the original BERT study. First, the input HS-CN is represented as $[CLS]\ HS\_tokens\ [SEP]\ CN\_tokens\ [SEP]$ and fed into BERT. By using the final hidden state of the first token [CLS] as the input, originally denoted as $C \in R^H$, we obtain a fixed-dimensional pooled representation of the input sequence. Then, a classification layer is added with the parameter matrix $W \in R^{KH}$, where K denotes the number of labels, i.e. 2 for HS-CN classification. The cross-entropy loss has been used during the fine-tuning.

We have conducted a hyperparameter tuning phase with a grid-search over the batch sizes 16 and 32, the learning rates [4,3,2,1]e-5 and the number of epochs in the range of 3 to 8. We obtained the best model by fine-tuning uncased BERT-large, with a learning rate of 1e-5, batch size of 16, and after 6 epochs at the 1029^{th} step on a single GPU.

The second model is built by fine-tuning ALBERT, which shows better performance than BERT on inter-sentence coherence prediction by using a sentence-order prediction loss instead of next-sentence prediction. In sentence-order prediction loss, while the positive examples are created similar to BERT by using the consecutive sentences within the same document, the negative examples are created by swapping sentences, which leads the model to capture the discourse-level coherence properties better (Lan et al., 2019). This objective is particularly suitable for HS-CN pair classification task, since HS and CN order and their coherence are crucial for our task. We fine-tuned ALBERT similarly to BERT model, by adding a classification layer on top of it. We applied the same grid-search that we used for BERT model to fine-tune ALBERT-xxlarge which contains 235M parameters. We saved a checkpoint at every 200 steps and finally, obtained the best model by using the learning rate of 1e-5, the batch size of 16, and at the 1200^{th} step.[4]

Metrics. To find the best model for machine reviewer, we compared BERT and ALBERT models over the test set. Although it seems more intuitive to focus on precision since we search for an effective filtering over many possible solutions, we observed that a model with a very high precision tends to overfit on generic responses, such as "*Evidence please?*". Therefore, we aim to keep the balance between the precision and recall and we opted for F1 score for model selection. We report the best configurations for each model in Table 6, and the percentage of filtered pairs in Table 5. ALBERT classifier outperformed BERT model in all three metrics; F1, Precision, and Recall. Considering 6% of absolute F1 score improvement with respect to BERT model, we employed ALBERT model as the Machine Reviewer.

Reviewer$_{machine}$	F1	Precision	Recall
ALBERT	0.73	0.74	0.73
BERT	0.67	0.69	0.65

Table 6: F1, Precision and Recall results for the two main classifier configurations we tested.

8 NGO Operators Experiments

To verify that the author-reviewer approach can boost HS-CN data collection, we run an experiment with 5 expert operators from an NGO. We compared the filtering strategies to reveal the best depending on several metrics.

[4] All the experiments have been conducted on a single GeForce RTX 2080 Ti GPU. Only the ALBERT classifier model has been trained with 8 TPU cores on Google Cloud.

Approach	NGO_{time}	$Crowd_{time}$	RR	Novelty	$Pairs_{selec}$	$Pairs_{final}$
no suggestion	480	-	2.72	-	-	-
$Reviewer_{expert}$	76	-	3.56	0.73	100%	45%
$Reviewer_{\geq 1}$	72	215	4.31	0.70	33%	54%
$Reviewer_{machine}$	68	-	4.48	0.68	40%	63%
$Reviewer_{\geq 2}$	49	703	5.70	0.65	10%	72%

Table 7: Results for CN collection under various configurations. RR for 'no suggestion' is computed on NICHE dataset and the time needed is the one reported in (Chung et al., 2019). Time is expressed in seconds. $Pairs_{selec}$ indicates the percentage of original author pairs that have been passed to the expert for reviewing, $Pairs_{final}$ indicates the percentage of selected pairs that have been accepted or modified by the expert themselves. $Crowd_{time}$ is computed considering that annotators gave a score every 35 seconds, and we required two judgments per pair.

Within Subject Design. We administered lists of HS-CN pairs to 5 operators from each filtering condition, and instructed them to evaluate/modify each pair in terms of 'suitableness' of the CN to the corresponding HS.

Instructions. For each HS-CN pair, we asked the operators: a) if the CN is a perfect answer, to validate it without any modification, b) if the CN is not perfect, but a good answer can be obtained with some editing, to modify it, c) if the CN is completely irrelevant and/or needs to be completely rewritten to fit the given HS, to discard it.

Measurement. The main goal of our effort is to reduce the time needed by experts to produce training data for automatic CN generation. Therefore the primary evaluation measure is the average time needed to obtain a proper pair. The other measurements of interest are Diversity and Novelty, to understand how the reviewing procedure can affect the variability of the obtained pairs.

Procedure and material. We gave the instructions along with a list of 20 HS-CN exemplar pairs for each condition (i.e. $Reviewer_{\geq 1}$, $_{\geq 2}$, $_{machine}$, $_{expert}$). The condition order was randomized to avoid primacy effect. In total, each NGO operator evaluated 80 pairs. Pairs were sampled from the pool of 2700 pairs described before (apart from the automatic filtering condition). To guarantee that the sample was representative of the corresponding condition, we performed a stratified sampling and avoided repeating pairs across subjects.

Results and Discussion. As it is shown in Table 7, there is a substantial decrease in data collection time (NGO_{time}) when automatic generation mechanisms are introduced (no suggestion vs. $Reviewer_{expert}$). If crowd filtering is applied ($Reviewer_{\geq 1}$, $_{\geq 2}$), the amount of time can be further reduced, and the more stringent the filtering criterion, the higher the time saved. Conversely, the more stringent the filtering criterion, the higher the time to obtain a filtered pair from non-expert annotators ($CROWD_{time}$). For instance to obtain a single pair with at least a score of 2 by both annotators, 700 sec (around 12 min) are needed on average (only 10% of examples are in ≥ 2 condition). Results indicate that providing an automatic generation tool meets the first goal of increasing efficiency of the operators in data collection.

Regarding diversity and novelty metrics, pre-filtering author's output ($Reviewer_{\geq 1}$, $_{\geq 2}$ and $_{machine}$) has a negative impact: the more stringent the filtering condition the higher the RR and the lower the novelty of the filtered CNs. We performed some manual analysis of the selected CNs and we observed that especially for the $Reviewer_{\geq 2}$ case (which was the most problematic in terms of RR and novelty) there was a significantly higher ratio of "generic" responses, such as "*This is not true.*" or "*How can you say this about an entire faith?*" , for which reviewers agreement is easier to attain. Therefore, the higher agreement on the generic CNs reveals itself as a negative impact in the diversity and novelty metrics. Conversely, the percentage of pre-filtered pairs that are accepted by the expert increases with the filtering condition becoming more stringent, the baseline being 45% for the $Reviewer_{expert}$ condition.

As for the amount of operators' effort, we observed a slight decrease in HTER[5] with the increase of pre-filtering conditions, indicating an improvement in the quality of candidates. However, HTER scores were all between 0.1 and 0.2, much below the 0.4 acceptability threshold defined by Turchi et al. (2013), indicating that operators modified CNs only if "easily" amendable.

[5]Human-targeted Translation Edit Rate is a measure of post-editing effort at sentence level translations (Specia and Farzindar, 2010).

Finally, we observe that despite reducing the ouput diversity and novelty, the reduction of expert effort by Reviewer$_{\geq 2}$ in terms of the percentage of the obtained pairs is not attainable by a machine yet. On the other hand, automatic filtering (Reviewer$_{machine}$) is a viable solution since (i) it helps the NGO operators save time better than human filter $\geq$1, (ii) it preserves diversity and novelty better than Reviewer$_{\geq 2}$ and in line with Reviewer$_{\geq 1}$.

9 Conclusions

To counter hatred online and avoid the undesired effects that come with content moderation, intervening in the discussion directly with textual responses is considered as a viable solution. In this scenario, automation strategies, such as natural language generation, are necessary to help NGO operators in their countering effort. However, these automation approaches are not mature yet, since they suffer from the lack of sufficient amount of quality data and tend to produce generic/repetitive responses. Considering the aforementioned limitations, we presented a study on how to reduce data collection effort, using a mix of several strategies. To effectively and efficiently obtain varied and novel data, we first propose the generation of silver counter-narratives – using large scale unsupervised language models – then a filtering stage by crowd-workers and finally an expert validation/post-editing. We also show promising results obtained by replacing crowd-filtering with an automatic classifier. As a final remark, we believe that the proposed framework can be useful for other NLG tasks such as paraphrase generation or text simplification.

Acknowledgments

This work was partly supported by the HATEMETER project within the EU Rights, Equality and Citizenship Programme 2014-2020. We are grateful to Stop Hate UK that provided us with the experts for the evaluation. Finally, there are also many people we would like to thank for their help and useful suggestions: Eneko Agirre, Simone Magnolini, Marco Turchi, Sara Tonelli and the anonymous reviewers among others.

References

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

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. *Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field- study/*.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Heiner Bielefeldt, Frank La Rue, and Githu Muigai. 2011. Ohchr expert workshops on the prohibition of incitement to national, racial or religious hatred. In *Expert workshop on the Americas*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Fabio Del Vigna, Andrea Cimino, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Ona de Gibert, Naiara Perez, Aitor Garcıa-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *EMNLP 2018*, page 11.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666.

Filip Klubička and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *LREC*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ruli Manurung, Graeme Ritchie, and Henry Thompson. 2008. An implementation of a flexible author-reviewer model of generation using genetic algorithms. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 272–281, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

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. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

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

Jon Oberlander and Chris Brew. 2000. Stochastic text generation. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1373–1387.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4757–4766, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690.

Tanya Silverman, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue, London. https://www. strategicdialogue. org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE. pdf–73*.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, pages 33–41.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Scott R Stroud and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, pages 293–310. Springer.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in mt quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

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*, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendices

HS	I love the Pres also he is a real man not a commie cunt like the last muslim in there.
GPT2	Use of the cword here is totally unacceptable because it is hurtful and offensive.
	That is very interesting, but they already have that kind of man who is always saying something different. No need for your comment.
	Hi there, please refrain from using derogatory comments in the thread. They are hurtful and unwanted. If you continue Admin will be alerted.
TRF	Please refrain from using derogatory terms for females.
	It's really hard to respect someone who uses racial slurs in an attempt at comedy.
	Please don't use hateful words, please use your words or else removal will take place.
HS	What cha doin America? Oh...just waiting to hang the cunt the Muslim nigga.
GPT2	Name calling speaks volumes about one's character.
	To act like those of a race are the same is to admit to your narrow mindedness.
	Hey there, racist and derogatory views are hurtful and uncalled for. Please refrain from this type of behavior or you will be reported. Thx.
TRF	It doesn't make sense to name call someone you've never met based on what the media says.
	I'd be more inclined to listen to your side if you didn't use that slur.
	Please muslims are not so two as a high everyone has no need to get an ignorant little elected official further.
HS	"Turkish Nazi Muslims" huh retard. Hey, try forming real sentences instead of half of your worlds being fucking hashtags, fuckwit.
GPT2	It is acceptable for you to talk like this but not okay to physically attack others with it.
	Some people find the Nazi Holocaust story appalling, but we should not use those slurs.
	Please refrain from using derogatory terms for the mentally handicapped.
TRF	Please do not use offensive language.
	Use of the r word is unacceptable because it hurts, insults and demeans people with mental disabilities.
	That word is offensive to the intellectually disabled. Furthermore, there is no shame in being intellectually disabled.
HS	The muslim nigger terrorist obama needs death the hussein needs it's limbs ripped off and his tongue seared with a hot poker.
GPT2	You don't know that, but I do.
	I understand your point but please watch the language.
	You are throwing a lot of hatred and anti-Semite in to this, which is not good.
TRF	I'd be inclined to agree with you if you didn't use slurs.
	Your reply isn't need, not big or clearer. Think before you type next time please.
	All religions, ethnicities, and social groups deserve tolerance.

Table 8: Randomly sampled CNs generated from GPT-2 and TRF models trained on CROWD dataset.

HS	White women converted to Islam are the saddest and most selfish persons that you can encounter. They praise happily Islam beliefs and become part of a sisterhood of tollerance and understanding, but they do not know the situation of women in Pakistan and other islamic states.
GPT2	Do you know that in many Muslim-majority countries, women's rights are recognized and actively promoted? Laws are passed that guarantee the right to have children, and foreign charities work to ensure that women have access to quality education and health services.
	Are you a believing woman who has known and respected great men like Mohammed, Albert and Winston Churchill. You have forgotten about those who sacrificed so much for your benefit.
	If white women who love Islam were doing such things why would their profile be so public?
TRF	Honestly, i see around many more muslim women without burqa or veil than women dressed like europeans.
	Are you sure? muslim nobel prize winners would suggest otherwise.
	Sounds like you are generalizing what a very small percentage of extremists do to an entire religion.
HS	We are every day in constant danger, because Muslims can blow us all up. For this fact we should put all Muslims in a hole in the ground.
GPT2	If you had said this about all the non-Muslim groups who may espouse hateful or violent views, how would you feel about being labelled a bigot?
	Apart from being offensive and ignorant, is also the belief that Muslims only create problems. In fact, many Muslims are working to solve some of these social problems.
	Muslims have never sought to destroy our world, rather they have offered their goods and services in return for peace and unity. Nostalgia may have something to offer us, but not for taking over countries and nations.
TRF	People who commit offenses are punished by law. Muslims are not all terrorists. muslims are not trying to generalize on their own community, indeed anything. If someone would like you, but you are peaceful muslim countries like which, please.
	No, a person can be immigrants for multiple allegiances and responsibilies.
	Why do you think that muslim criminals will receive different treatment or even privileges?
HS	Koran is a book that should be burnt and mosques should be demolished. If we do all these things, all problems in our community will be solved.
GPT2	All book burning is prohibited under the Universal Declaration of Human Rights, which says: ' Everyone has the right to book his escape from cruelty and oppression wherever it may be found.'?
	Yes, because this book encourages people to respect for life and to respect God. Glad you like it!
	Praise be to God that we have these humble requests. Can you give us some figures on the number of people who would like to migrate to your country if you could only send em home?
TRF	If you dislike diversity so much , why do not you go off and live in a country where there are no muslims, only people like you?
	Mgm fgm have been in britain for hundreds of years. it is never acceptable but it does not come from just one country or ideology.
	The idea of blow someone up is itself crazy.

Table 9: Randomly sampled CNs generated from GPT-2 and TRF models trained on NICHE dataset.

HS	Girls and boys are brainwashed by the same people.
No filter	That's why most religions end well.
Reviewer$_{\geq 1}$	Mindless Hate and fear mongering divide and conquer.
Reviewer$_{\geq 2}$	Mindless Hate and fear mongering do not help anyone.
HS	Everything will be better if Muslims did not exist. Their only contribute to our society is violence.
No filter	Did you know that the greatest child abusers in British history were white men? That is if they weren't SO EXTREMELY LIE!?!?!
Reviewer$_{\geq 1}$	Why would you say that? I would say that all the people using social media and whatever to vent their anger and frustration are clearly not serious about their religion. Some may be, and are angry, but that does not mean they are responsible for anything.
	The world would be a better place without forced baptisms and child rapists.
	What exactly are you suggesting?
Reviewer$_{\geq 2}$	The world would actually be a very dark place without Muslims who contribute a lot to our society. How about a place of peace, tolerance and understanding for all?
HS	Europe is civilised, Muslims should not stay there.
No filter	That is a misunderstanding. On the one hand, Europe is a civilized place, on the other, Muslims cannot excel at anything.
Reviewer$_{\geq 2}$	How can you say that about an entire faith of 1.6 billion people?
Reviewer$_{machine}$	Muslims have been a part of the European society for centuries, they have been a part of its institutions since the middle ages, they have been active in politics during the 19th century and have been crucial in the development of Europe during the last decades.
HS	Multiculturalism has brought us nothing but disaster.
Reviewer$_{machine}$	The multiethnic society has produced many smart and talented people, who have gone on to work in prominent public offices. No one is asking us to sacrifice our culture or our identity for the greater good of the community.

Table 10: Randomly sampled CNs, generated from GPT-2, filtered from various reviewer configurations.