arXiv:2211.03433v1 [cs.CL] 7 Nov 2022

Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering

Helena Bonaldi[1,3], **Sara Dellantonio**[2,3], **Serra Sinem Tekiroğlu**[3], **Marco Guerini**[3],
[1]University of Trento, Italy
[2]Free University of Bozen-Bolzano, Italy
[3]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
hbonaldi@fbk.eu, sdellantonio@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu

Abstract

Fighting online hate speech is a challenge that is usually addressed using Natural Language Processing via automatic detection and removal of hate content. Besides this approach, counter narratives have emerged as an effective tool employed by NGOs to respond to online hate on social media platforms. For this reason, Natural Language Generation is currently being studied as a way to automatize counter narrative writing. However, the existing resources necessary to train NLG models are limited to 2-turn interactions (a hate speech and a counter narrative as response), while in real life, interactions can consist of multiple turns. In this paper, we present a hybrid approach for dialogical data collection, which combines the intervention of human expert annotators over machine generated dialogues obtained using 19 different configurations. The result of this work is DIALOCONAN, the first dataset comprising over 3000 fictitious multi-turn dialogues between a hater and an NGO operator, covering 6 targets of hate.

1 Introduction

While hate towards vulnerable groups or individuals is not a new phenomenon, the upsurge of hate speech and its proliferation is relatively recent and it is enabled by the fast spread of information in online platforms. The rise in hate speech online can even provoke violent actions offline. Consequently, fighting online Hate Speech (HS) has become a vitally important "job for everyone"[1] especially for the NLP researchers. The contrast to HS and haters on social media platforms is usually carried on via user suspension, content removal or shadow banning, which can be mapped to a classification task in NLP terms. However, AI and NLP can play even a more crucial role that is not limited to classification. In fact, recently, NLG models have started to be proposed as an effective tool to counter HS by providing relevant responses. In particular, the idea is to imitate the operators of Non-Governmental Organizations (NGO) that are actually intervening in online discussions by replying to hateful content using so-called Counter Narratives (CN), defined by Schieb and Preuss (2016) as "communicative actions aimed at refuting hate speech through thoughtful and cogent reasons, and true and fact-bound arguments". Through automatically generating CNs, it is possible to aid NGO operators in their day-to-day manual activities, and therefore to partially countervail the sheer amount of hateful content posted online (Chung et al., 2021b).

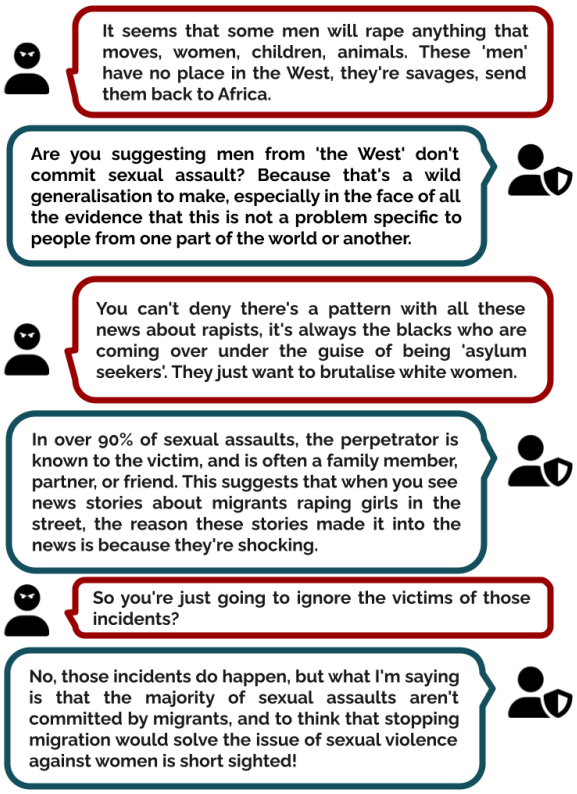

Figure 1: Exemplar dialogue between a hater and an NGO operator.

Despite the invaluable attempts to create HS/CN

[1]"Hatred is a danger to everyone – and so fighting it must be a job for everyone." *António Guterres, United Nations Secretary-General, 2021*

datasets and systems (Mathew et al., 2019; Qian et al., 2019; Chung et al., 2019; Fanton et al., 2021), up to now only datasets containing 2-turn interactions have been proposed (i. e. a hate speech and a responding counter narrative), while in real scenarios, such as on social media platforms, multi-turn dialogues are the norm. In Figure 1 an example of such dialogues is provided[2]. Therefore, multi-turn dialogue datasets are necessary for training models that can better handle online hate phenomenon.

Still, obtaining expert-written quality data to train such models on is not trivial. To ameliorate this problem, a recently proposed approach is the use of *hybrid* data collection strategies where a human and a machine collaborate to build data starting from a seed dataset of expert based examples (Fanton et al., 2021). In this paper we follow this line of research and investigate novel strategies and algorithms that are specifically designed for multi-turn dialogues collection.

In particular, we test 19 different *hybrid* strategies obtaining a novel dataset of more than 3K dialogical interactions between two interlocutors, one acting as the hater and the other as the NGO operator, for a total of more than 16K turns. We call this dataset DIALOCONAN (DIALOgical COunter-NArratives collectioN). This is the first and most comprehensive multi-target dataset that addresses expert-based counter narrative generation in fully dialogical scenarios, and it can be downloaded at the following link: `https://github.com/marcoguerini/CONAN`.

2 Related Work

In this work, we consider four main research areas as relevant: in particular (i) available datasets for hate speech detection, (ii) available datasets for CN generation, (iii) CN generation approaches, and (iv) hybrid data collection methodologies.

Hate detection. Many benchmarks for automatic HS detection are currently available (Mathew et al., 2021; Cao et al., 2020; Kumar et al., 2018; Hosseinmardi et al., 2015; Waseem, 2016; Burnap and Williams, 2016). Regarding the systems built on top of these benchmarks, we refer the readers to the surveys by Poletto et al. (2020); Schmidt and Wiegand (2017); Fortuna and Nunes (2018) for detailed reviews. Other reviews include the analysis of ethical implications (Kiritchenko et al., 2021) and of problems such as bias replication (Binns et al., 2017; Davidson et al., 2019; Vidgen and Derczynski, 2020; Sap et al., 2019; Tsvetkov, 2020).

CN data collection. Since CNs have been shown to be effective in reducing linguistic violence (Benesch, 2014; Gagliardone et al., 2015; Schieb and Preuss, 2016; Silverman et al., 2016; Mathew et al., 2019) and in changing the viewpoints of bystanders (Allison and Bussey, 2016; Anderson et al., 2014), they are beginning to be collected as training data for supervised NLG models. The investigated approaches for data collection can be listed as crawling (Mathew et al., 2018, 2019; Yu et al., 2022), crowdsourcing (Qian et al., 2019), nichesourcing (Chung et al., 2019) and hybrid approaches (Tekiroğlu et al., 2020; Fanton et al., 2021).

The most relevant datasets for our work are (i) Fanton et al. (2021) in terms of quality and target diversity, even if it only includes HS/CN pairs, and (ii) Qian et al. (2019) that hints at the issue of multi-turn dialogues. However, in the latter the CN is only the last turn of a forum-style dialogue among more than 2 interlocutors, rather than a HS/CN multi-turn dialogue between two opposing actors.

CN generation. Neural approaches to generate CNs have started to be studied along with available datasets (Fanton et al., 2021; Tekiroğlu et al., 2020; Qian et al., 2019). Tekiroglu et al. (2022) present a thorough comparison of several pre-trained LMs for this task. Zhu and Bhat (2021) propose an entirely automated 2 stage pipeline where several CN candidates are generated and then filtered. Other lines of work include CN generation for under-resourced languages (Chung et al., 2020), or the generation of knowledge-bound CNs, to avoid hallucination phenomena (Chung et al., 2021a). Finally, Ashida and Komachi (2022) studied CN generation with LLMs, using few-shots prompting.

Hybrid models for data collection. A recently emerged data collection methodology is based on *hybrid* models, where humans and machines work together to collect better quality data in a more efficient way. Wallace et al. (2019) propose using model output to guide humans in the writing of adversarial examples for question-answering systems. Dinan et al. (2019) and Vidgen et al. (2020) perform a data collection for offensive language detection with repeated model-human interactions where the classifier output drives annotators in ex-

[2]This paper includes examples of hateful content, which may be upsetting for the readers. However, they do not represent the views of the authors.

ample creation at each round. A more recent study proposes a *hybrid* approach where an LM is trained to generate HS/CN pairs that are validated and post-edited by annotators (Tekiroğlu et al., 2020). Fanton et al. (2021) further expand this approach by making it iterative using several LM configurations.

3 Methodology

Data collection can be very difficult and time consuming when high quality data from experts are necessary. Given that we need to collect whole HS/CN dialogues and not just pairs, the problem is even harder. Moreover, scraping NGO operators' real interactions is not a viable solution, considering that this data can be used for account "doxing". In fact, malicious users could reverse-search the text included in a dataset to identify the operators' accounts. This would undermine their work, since they usually operate undercover, and would expose them to possible attacks.

Therefore, we decided to resort to *hybrid* approaches and run 3 different data collection sessions based on the aspects of the dialogue augmentation we want to address (either the structure, in terms of turns order, or the wording of the turns).

In total we tested 19 different dialogue collection strategies. All the strategies are inserted in an author-reviewer pipeline as described by Tekiroğlu et al. (2020), where the *author* is a single dialogue creation strategy at a time, and the *reviewer* is represented by a team of trained annotators, who are tasked with post-editing the dialogues generated by the given author strategy.

Author - Configurations. Each of the 3 data collection sessions we perform has different input data and author tasks, in particular:

- **Session 1: same wording, new dialogue structure.** 7 strategies based on concatenating pre-existing material (HS/CN pairs) to obtain new dialogues.
- **Session 2: new wording, same dialogue structure.** 6 strategies to modify the wording of pre-existing dialogues via paraphrasing.
- **Session 3: new wording, new dialogue structure.** 6 strategies using generative Language Models (LMs) for complete dialogue generation.

Author - Seed datasets. Since each author configuration needs some textual input, we employ (i) a dataset, created ad hoc, consisting of 222 fictitious dialogues and (ii) HS/CN pairs coming from the dataset presented in Fanton et al. (2021).

The ad hoc fictitious dialogues (DIALO$_{gold}$ henceforth) are written by two expert NGO operators, who have been working for over 10 years in writing CNs on social media platforms. They were asked to write dialogues between a hypothetical hater and an NGO operator, following their real expertise in the task. The dialogues can have 4, 6, or 8 turns (these are typical lengths according to their experience) and cover the following 6 targets of hate, defined beforehand: `LGBT+`, `MIGRANTS`, `MUSLIMS`, `JEWS`, `POC` and `WOMEN`.

Given the small size of DIALO$_{gold}$, we also use part of the dataset presented in Fanton et al. (2021) as an additional resource. This dataset consists of 5000 HS/CN pairs covering, among others, the 6 targets of hate present in DIALO$_{gold}$. Therefore, we extracted the pairs labeled with these 6 targets so that the two resources can be 'aligned' by topic, and we named it PAIRS$_{gold}$, since also this dataset was created with the help of expert NGO operators.

Reviewers - Training. For post-editing the output of the various author configurations, three annotators were recruited from a pool of internship students. They have been extensively trained using the methodology of Fanton et al. (2021), in order to become "experts" on HS/CN post-editing. In particular, we first explained the aim of the task. Then, they had to read NGO guidelines and documentation on CN writing[3], together with all the dialogues present in DIALO$_{gold}$, which were provided as examples of the material they would have to work with. We detailed the methodology, explaining that the main focus was to make the dialogues natural, with the minimum intervention possible and keeping the seed dataset as a reference for naturalness. General instructions about the post-editing procedure were also provided, pointing out that for each session specific guidelines would have been given.

Reviewers - Mitigation procedure. Finally, we also implemented a mitigation procedure similar to the one presented by Vidgen et al. (2019). This procedure is implemented to safeguard the annotators' well-being while working with abusive content and it includes: (i) explaining to the annotators the pro-social nature of the research and the purpose of

[3]See `https://getthetrollsout.org/stoppinghate` as a reference.

their post-editing activity, (ii) advising the annotators to work few hours per day and to take regular breaks (iii) having weekly meetings to let possible problems or distress emerge.

Data collection procedure. For each session we applied the following procedure: (i) generate dialogue candidates according to session specific strategies, (ii) adapt the annotation guidelines to the specific session, (iii) let the annotators practice the task on a small "training" set of dialogue candidates, and (iv) update the guidelines with respect to their feedback. Lastly, (iv) annotators complete the post-editing on the remaining dialogues following the updated guidelines (the order of the dialogues was randomized to avoid comparison or primacy/recency effects over session strategies).

4 Metrics

We use several metrics to assess the performance of each strategy. These metrics are aimed to assess either the *efficiency* of the procedure or the *quality* of the obtained data.

HTER is an efficiency metric used to measure the post-editing effort of the annotator, and it is usually employed for sentence level translations (Specia and Farzindar, 2010). A value above 0.4 is generally used to account for low quality outputs, where rewriting from scratch is on par with correcting it (Turchi et al., 2013).

Turn deletion is the percentage of turns that are discarded by the reviewers since their quality is too low and/or they do not fit in the current dialogue structure. The more content needs to be deleted, the less efficient the procedure is.

Turns swap is the percentage of turns that are moved by the reviewers from the original position they were in, to another position in the final edited dialogue. Usually turns of this kind have a good quality but they do not fit the current position.

Novelty is utilized to check the quality of a generated dialogue by measuring its lexical difference with respect to a reference set of dialogues, and it is grounded on Jaccard similarity (Dziri et al., 2019; Wang and Wan, 2018).

Repetition Rate (RR) measures the language diversity within a corpus using the rate of non-singleton ngram types (Cettolo et al., 2014; Bertoldi et al., 2013). It is used in our experiments to evaluate each strategy in terms of its ability to provide diverse and varied examples.

5 Session 1: Dialogue structure

In Session 1 we started from the HS/CN pairs in PAIRS$_{gold}$ and concatenated them in order to produce dialogue candidates with different structures.

5.1 Author Strategies

We employ 7 strategies to connect HS/CN pairs from PAIRS$_{gold}$ to create 4, 6, and 8 turns examples (consistently with the DIALO$_{gold}$ characteristics).

During the concatenation, each pair is used only once in a dialogue. The connection strategies are: random concatenation (1 strategy), similarity concatenation (4 strategies), and keyword matching concatenation (2 strategies). In order to obtain a balanced dataset, for each strategy, each target, and each dialogue length combination we created 10 connected dialogues. In-detail descriptions of the 7 concatenation strategies we utilized are as follows:

Random connection. For the random connection (**RND**), the selected pairs for each target are randomly concatenated to form dialogues. This strategy represents a baseline to which we compare against while analysing the other strategies.

Similarity connection. To connect pairs depending on to their similarity, we utilize (i) the Jaccard similarity and (ii) the cosine similarity[4]. Both for the Jaccard and cosine similarity, we perform pair matching via two approaches to form the $HS_i, CN_i, HS_{i+1}, CN_{i+1}$ concatenation:

1. **SIM**$_{HS\text{-}HS}$ = the similarity between HS_i and HS_{i+1};
2. **SIM**$_{CN\text{-}HS}$ = the similarity between CN_i and HS_{i+1};

For each pair, we randomly select 1 among the 10 most similar pairs according to the chosen similarity (either Jaccard or cosine) and concatenation elements (either HS-HS or CN-HS). The procedure is repeated until the desired number of turns for each dialogue is reached.

Keywords connection. We employ the YAKE keyword extractor (Campos et al., 2020) to extract two keywords from each HS and CN of PAIRS$_{gold}$ and perform a concatenation similar to the previous strategies. We connect HS_i, CN_i and HS_{i+1}, CN_{i+1} according to the following criteria:

[4]Cosine Similarity is computed on their embeddings obtained with `mpnet-base`. The Sentence Transformer library (`https://www.sbert.net/`) has been employed.

	Efficiency			Quality			
	del turns	**HTER**	**swap**	**RR**$_{gen}$	**RR**$_{ed}$	**NOV**$_{g\text{-}g}$	**NOV**$_{g\text{-}e}$
RND	12.222	**0.141**	20.926	**4.286**	**4.482**	0.820	0.818
J-SIM$_{HS\text{-}HS}$	14.259	0.193	**15.185**	9.353	5.964	0.827	0.823
C-SIM$_{HS\text{-}HS}$	10.370	0.186	15.926	9.450	5.215	0.824	0.820
KW$_{HS\text{-}HS}$	21.111	0.283	**14.444**	15.774	4.710	**0.828**	0.823
J-SIM$_{CN\text{-}HS}$	10.741	0.145	23.333	7.454	6.204	0.826	**0.824**
C-SIM$_{CN\text{-}HS}$	**8.889**	**0.134**	18.704	**7.128**	**5.087**	0.820	0.818
KW$_{CN\text{-}HS}$	**8.197**	0.152	15.222	11.710	9.035	**0.828**	**0.824**

Table 1: Results for the first session. J-SIM and C-SIM are the connections via Jaccard and cosine similarity, respectively. RR$_{gen}$ and RR$_{ed}$ are respectively the RR of the data before and after post-editing, while NOV$_{g\text{-}g}$ and NOV$_{g\text{-}e}$ are the novelty of the data before and after post-editing with respect to DIALO$_{gold}$.

1. **KW**$_{HS\text{-}HS}$ = if HS_i and HS_{i+1} share two keywords;

2. **KW**$_{CN\text{-}HS}$ = if CN_i and HS_{i+1} share two keywords;

We decided on a 2-keywords match since according to our preliminary manual analysis we found that the first keyword is often target-related; by considering two keywords we aim to include also a topic-related keyword.

As a final note, we should highlight that the two groups of connection strategies (HS-HS and CN-HS) represent either (i) a *global* semantic coherence across turns (all HS being similar) or (ii) a *local* semantic coherence (only CN-HS of adjacent turns being similar) both for SIM and KW. By using a *global* semantic coherence via HS-HS matching we attempted to simulate the attitude of the hater which is convinced of their own ideas and do not accept any external input, while with the *local* connections, we aimed to recreate a "linguistic alignment" phenomenon (Doyle and Frank, 2016). Details on the matching procedures and the description of the algorithms for SIM and KW we employed are reported in Appendix A.1.

5.2 Reviewing phase and guidelines

In order to obtain natural dialogues, the annotators in this session received specific post-editing instructions:

1. Since CNs are gold, it is strongly suggested to post-edit only the HS$_{i+1}$ to "align" it with the CN$_i$ belonging to the previous turn.

2. If a pair is in an unnatural position of the dialogue it should be moved to a better position.

3. If a pair is not fitting with the flow of the dialogue and cannot be moved elsewhere, it should be deleted.

4. If the whole dialogue makes no sense, or is too difficult to fix, it should be deleted.

A characteristic example of the post-editing done in Session 1 is shown in Table 10 in Appendix C.

5.3 Results

Results of this session in terms of *efficiency* and *quality* are reported in Table 1. In general, we observe that strategies using any HS-HS connection are less efficient, having higher HTER scores as compared to the CN-HS ones. HS-HS connections also have a high rate of deleted turns, in particular KW$_{HS\text{-}HS}$ and J-SIM$_{HS\text{-}HS}$. The KW$_{HS\text{-}HS}$ strategy is even more inefficient than the random connection baseline (it reaches the highest number of deleted turns and the highest HTER), and it is the most repetitive before post-editing, as showed by the RR$_{gen}$. These results are also confirmed by the annotators' feedback, who noted the presence of dialogues which were particularly difficult to edit since they contained the same HS repeated multiple times (see example in Table 11, Appendix C). A *posteriori* analysis showed that these dialogues were mainly obtained through the KW$_{HS\text{-}HS}$ connection. Moreover, each HS-HS connection strategy achieves a higher RR$_{gen}$ score than its CN-HS counterpart, showing that connecting through a *global* similarity generates a higher overall repetitiveness than using a *local* similarity. The particular high scores reached by the RR$_{gen}$ of both the keywords connection strategies can be explained by the procedure employed for connection: for keywords we performed an exact matching, whereas with the cosine or Jaccard similarity, the connection was selected from the 10 most similar candidates.

After post-editing, all the strategies achieve a lower RR, between 4.5 and 9, indicating a more diversified content. The novelty is calculated against DIALO$_{gold}$: the scores are similar for all the strate-

	Efficiency	Quality					
	HTER	**RR**$_{gen}$	**RR**$_{ed}$	**NOV**$_{g\text{-}g}$	**NOV**$_{g\text{-}e}$	**NOV**$_{mt\text{-}g}$	**NOV**$_{mt\text{-}e}$
Neutral$_1$	0.355	4.019	3.684	0.749	0.770	0.258	0.450
Neutral$_2$	0.398	**3.943**	**3.275**	**0.775**	**0.774**	**0.470**	**0.472**
Style$_{tw}$	0.355	**3.836**	**3.396**	0.756	0.773	0.327	0.465
Style$_{dialo}$	**0.348**	5.452	4.253	0.743	0.765	0.388	0.465
Style$_{formal}$	**0.332**	4.512	3.710	0.751	0.764	0.359	0.450
Style$_{casual}$	0.369	4.346	4.118	**0.763**	**0.774**	**0.416**	**0.468**

Table 2: Results for the second session. The showed paraphrasers are, from top to bottom: the Protaugment and Style paraphraser with basic, Twitter and Switchboard style, and the Style former paraphrasers with formal and casual style. NOV$_{mt\text{-}g}$ and NOV$_{mt\text{-}e}$ are the novelty of the generated and post-edited data with respect to the dialogues resulting from Session 1.

gies, and they are hardly affected by the post-editing, showing that each strategy managed to add a consistent novelty to the already present gold data. Finally, it is worth noting that the strategies employing HS-HS connections have less turn swaps than C-SIM$_{CN\text{-}HS}$ and J-SIM$_{CN\text{-}HS}$. The most probable explanation is that CN-HS strategies require less deletion, but this comes at the cost of more turn swaps.

6 Session 2: Dialogue Wording

In the second session we focused on strategies aiming to obtain a new wording, given a structured dialogue. In particular, we tested 6 paraphrasing approaches on DIALO*gold* and on a part of the dialogues resulting from the first session. In this session, our overall aim is to obtain novel and diverse responses to hate. Therefore, we chose to paraphrase only the CNs belonging to a subset of the data collected in Session 1, while keeping the corresponding HS as it is.

6.1 Author Strategies

We carried out two exploratory studies to test different paraphrasing configurations and we selected the 6 most promising ones, as described in Appendix A.2. We use both paraphrasers with no specific style and with style transfer in order to attain a diverse data collection. The selection has been performed by assessing the aspects of dialogue wording that we deem the most relevant for our scenario.

Basic paraphrasing. We use 2 paraphrasing tools as a 'baseline' where we do not impose any specific style to the paraphrases: the Protaugment paraphraser (Dopierre et al., 2021) and the Style paraphraser (Krishna et al., 2020) with basic style.

Style paraphrasing. This group includes 4 strategies in which we aimed to generate paraphrases with specific styles, in order to enhance the diversity of our data collection. Specifically, we focused on a style similar to that present in social media or in dialogues (Style paraphraser (Krishna et al., 2020) with Twitter and Switchboard style), and formal or casual (Style former paraphraser[5] with casual and formal style).

For each CN, 3 different paraphrases are generated using the same paraphrasing strategy.

6.2 Reviewing phase and guidelines

In order to obtain more natural examples, the post-editing instructions given to the annotators are adapted accordingly, emphasizing the significance of novel wording.

1. The annotator should keep the gold HS as it is, while post-editing the most promising among the 3 CN paraphrasis suggestions, i. e. the one introducing the least errors and the most different one from the original.

2. Turn swap in this case is not allowed, since turns order was already validated in these dialogues and paraphrasing would not affect it.

3. For the same reason, turn and dialogue deletion are not allowed.

An example of a typical intervention of the annotators in Session 2 is shown in Table 12 in Appendix C.

6.3 Results

We report the results in terms of *efficiency* and *quality* in Table 2. All the paraphrasers employed

[5] https://github.com/PrithivirajDamodaran/Styleformer

	Efficiency			Quality					
	del turns	**HTER**	**swap**	**RR**$_{gen}$	**RR**$_{ed}$	**NOV**$_{t\text{-}g}$	**NOV**$_{t\text{-}e}$	**NOV**$_{g\text{-}g}$	**NOV**$_{g\text{-}e}$
DGPT$_{b}$	50.179	0.678	8.214	**7.815**	**3.976**	**0.793**	0.804	0.787	0.793
DGPT$_{mt}$	**16.786**	0.408	10.714	**8.587**	6.110	0.757	0.759	0.798	0.799
T5$_{b\text{-}1m}$	76.875	0.655	0	16.672	**5.651**	0.789	**0.817**	0.783	**0.804**
T5$_{mt\text{-}1m}$	**34.375**	**0.362**	0	10.605	6.950	0.756	0.756	**0.802**	**0.803**
T5$_{b\text{-}2m}$	85.000	0.603	0	20.658	5.764	**0.804**	**0.805**	0.793	0.794
T5$_{mt\text{-}2m}$	38.929	**0.376**	0	10.756	7.678	0.756	0.756	**0.799**	**0.803**

Table 3: Results for the third session: the baseline models are signaled by the subscript $_b$, while the models trained on both DIALO$_{gold}$ and the dialgues resulting from Session 1 have the subscript $_{mt}$. NOV$_{t\text{-}g}$ and NOV$_{t\text{-}e}$ are respectively the novelty scores of the generated and post-edited data with respect to each model's training data.

reach similar HTER scores, which are below the 0.4 threshold, but higher than Session 1 results. Regarding the quality, generated paraphrases are highly novel with respect to the dialogues present in DIALO$_{gold}$, but not as high if compared to the dialogues resulting from the connection of the gold pairs in Session 1. In addition, the annotators' intervention enhances the novelty of the generated paraphrases in almost all the cases, and reduces the RR for all the paraphrasers, with lower scores than in the first session (3,739 vs. 5,814 on average).

To sum up, we conclude that it is better to concatenate PAIRS$_{gold}$ if we have a high number of pairs available, while paraphrasing is a viable solution if there is no pairs availability, since it implies a higher HTER and it is not justified by higher novelty.

7 Session 3: Generation

In this session we follow the overall configuration presented in Fanton et al. (2021), where the author is an LM fine-tuned on the DIALO$_{gold}$ together with the dialogues resulting from Session 1[6].

7.1 Author Strategies

We tested the following configurations:

DialoGPT. An autoregressive model specific for dialogue generation (Zhang et al., 2020). We choose DialoGPT since it is proven to be effective in CN generation as well (Tekiroglu et al., 2022);

T5$_{2m}$**.** Two T5 (Raffel et al., 2020) models conversing with each other: one fine-tuned to produce only HS and one to produce CNs. This configuration allows to completely decouple CN production from HS production.

T5$_{1m}$**.** One T5 model able to produce both HS and CN. We test it as a comparison to the two T5 models conversing with each other.

For each configuration, we test a baseline model, fine-tuned on DIALO$_{gold}$ only, and a model fine-tuned on both DIALO$_{gold}$ and the post-edited dialogues resulting from Session 1. For each model we employed the Top-p decoding mechanism (Holtzman et al., 2020) with $p = 0.9$[7]. In all cases, we split the employed dataset into training, development, and test sets with a ratio of 8:1:1. For the generation phase, we use as a prompt the initial HS of the test set dialogues. Then, we generate a single turn at a time by feeding the model with the context generated so far, until we reach 8 turns dialogues.

7.2 Reviewing phase and guidelines

The data generated with the LMs include both HS and CN, therefore the annotators are allowed to post-edit both, unlike the previous session. The reviewing guidelines are similar to those for Session 1, with the following changes:

1. it is possible to swap single turns and not only pairs, since the connection between HS/CN is not granted a-priori as in the previous sessions[8].

2. if some turns in a dialogue have a clearly different target than the labeled one, they should try to change turns wording to fit the original target.

3. the annotators should check the veracity of fact-based statements since they might derive from LM hallucinations.

[6] We did not include the dialogues resulting from Session 2 since they would have added little novelty. In particular, Session 2 CNs are paraphrases of those present in Session 1, and the HS of the dialogues in the two sessions are identical.

[7] Training details are reported in Appendix B.

[8] For example, a model can introduce hateful content when it is supposed to generate a CN, or viceversa (as showed in Table 13, Appendix C).

	Efficiency			Quality				Syntactic Complexity				
	del turns	**HTER**	**swap**	**RR**$_{gen}$	**RR**$_{ed}$	**NOV**$_{g\text{-}g}$	**NOV**$_{g\text{-}e}$	**avg turn len**	**avg turn #**	**MSD**	**ASD**	**NST**
Gold	-	-	-	-	-	-	-	**25.873**	5.105	**5.515**	**4.722**	**1.785**
Session 1	**12.381**	**0.175**	**17.753**	**9.112**	**5.382**	**0.824**	**0.821**	**20.065**	**5.833**	**4.828**	**4.236**	1.629
Session 2	-	**0.360**	-	**4.244**	**3.611**	0.756	0.770	19.6108	5.705	4.778	**4.236**	1.578
Session 3	**40.801**	0.448	**2.07**	10.646	6.385	**0.795**	**0.800**	19.944	**6.172**	4.757	4.112	**1.655**

Table 4: Results of the data collected at each session.

7.3 Results

Results of this session, in terms of *efficiency* and *quality*, are reported in Table 3. There are two major conclusions we can draw[9].

Firstly, adding the post-edited dialogues obtained concatenating PAIRS$_{gold}$ to the training data (DIALO$_{gold}$) strongly increases the efficiency. In fact, these models require much less deletion from the annotators with respect to the baselines, reaching a lower HTER ($<= 0.4$). Also, even if the dialogues generated with the baselines have a higher novelty with respect to the training data, they are also extremely repetitive in almost all cases.

Secondly, as already shown by Tekiroglu et al. (2022), autoregressive models are producing more varied and relevant content as compared to seq2seq models. In fact, even if DialoGPT requires more post-editing than T5 configurations (with comparatively higher HTER scores), its output dialogues require a lower number of deletion. This indicates that that the DialoGPT generation is suboptimal but rarely unsuitable, while this often is not achieved by T5. In particular, turns swaps are present only for the DialoGPT models. According to the annotators, this is explained by the characteristics of T5 dialogues, which are more stereotypical, vague but have a better structure (see Table 14 in Appendix C). This is also confirmed by the quality results: T5 models generate content with similar novelty scores to DialoGPT, but they also tend to be more repetitive.

8 Session comparison & Data description

Finally, Table 4 compares the results for each session over the main metrics of interest. We observe that concatenating pre-existing material that is already verified (i.e. HS/CN from PAIRS$_{gold}$) requires less effort than generating new data from scratch or paraphrasing gold material, as Session 1 reaches a lower HTER than both Session 2 and Session 3. On the other hand, in terms of the structure of the dialogues, Session 1 requires the highest effort as shown by the high swap rate. Meanwhile, Session 2 is the least repetitive, but also the least novel, providing dialogues with a good wording, even if this is not accompanied with a novel content. In general, all the sessions reach an HTER lower or equal to 0.4, and similar novelty scores with respect to the gold data. Therefore, in all cases it was possible to enhance the novelty of the initial seed dataset, with a reasonable post-editing effort.

Table 4 also shows a syntactic analysis of the data collected with each session, calculated at turn-level. The dialogues generated with the Language Models achieve the most balanced distribution in terms of number of turns[10], at the cost of simpler turns, as shown by the low maximum syntactic depth (MSD) and average syntactic depth (ASD) reached by Session 3. Paraphrasing instead provides the shortest generations both in terms of average turns length and of number of sentences (NST).

By comparing the results of the different sessions, we can conclude that the choice of the preferable data collection strategy firstly depends on the available input data, e. g. we might not always have gold HS/CN pairs available or multiple turns dialogues. Secondly, depending on the desired output, if the priority is to obtain novel content, Session 2 strategies would be the least favorite. Also, the concatenation of existing pairs as in Session 1 is a more cautious approach than the generation of completely new dialogues through LMs. Thus, Session 1 strategies can be preferred for a more conservative approach, whereas Session 3 strategies are better suited for a more creative data collection that comes at the cost of higher human correction effort.

As a last step, we performed a sanity check in which a senior NGO expert conducted a qualitative evaluation by reading a random sample

[9]While our main focus is on dataset creation, the results of this session offer also a form of simple benchmarking and some useful insights for the development of new models. In fact, the various metrics that we employed (post-editing, turn deletion, etc.) already provide a good indication of the LMs performance, especially for an open-ended scenario.

[10]We collected dialogues with 4, 6 and 8 turns, so a perfect balance would be of 6 turns.

	Dialogues	Coverage
JEWS	468	15.30
LGBT+	591	19.32
MIGRANTS	534	17.46
MUSLIMS	505	16.51
POC	493	16.12
WOMEN	462	15.10
Other	6	0.20
Total	3059	100

Table 5: The distribution of targets in the final dataset. 'Other' indicates a few cases of intersectional targets among the 6 given, e.g. MUSLIMS/WOMEN.

of the post-edited dialogues from each session. Their feedback was positive, no critical issues were raised and all the dialogues were approved both in terms of produced CNs and of their overall structure/naturalness. Our final dataset, DIALOCONAN, includes also the dialogues collected through the various training phases and exploratory studies of the annotators. Table 5 shows the distribution of targets in terms of number and percentage of dialogues. The distribution is reasonably balanced, with the LGBT+ target being the most represented. Overall, we collected 3059 dialogues for a total of 16625 turns.

9 Conclusion

In this paper we have presented a hybrid approach for dialogue data collection in the realm of hate speech countering. These dialogues have been obtained starting from two expert-based seed datasets and then combining the intervention of human annotators over machine generated dialogues. We tested 19 different strategies for generation, focusing on two crucial aspects of dialogue, i.e. structure and wording. We analysed all these strategies in terms of efficiency of the procedure and quality of the data obtained. The result of this work is DIALOCONAN, the first dataset comprising over 3000 fictitious multi-turn dialogues between a hater and an NGO operator, covering 6 targets of hate.

Acknowledgements

We are deeply thankful to Stop Hate UK and its volunteers for the help in writing the dialogues for the DIALO$_{gold}$ seed dataset and for sharing their expertise, fundamental to this work.

Limitations

The datasets currently available for CN generation are mainly for the English language and this one is no exception. The problem is that getting in contact with NGO operators for other languages is not easily solvable. The alternatives, such as translating this dataset into other languages represent a suboptimal solution. In fact each language and country has (i) its own peculiar canards against minorities, (ii) even if HS can be ported across countries (e.g. "*Migrants steal our jobs.*"), the arguments to counter such HS may vary (e.g. different laws, different socioeconomic situations, different statistical data). When translating dialogues all these nuances get lost reducing the possible effectiveness and introducing possible unnatural answers making reference to the country for which the original CN was written.

Although we tried to keep the overall quality of the final output as high as possible, since the dataset is created through a human-machine collaboration paradigm, it can still not be on par with the data that we can potentially obtain with niche sourcing the whole dataset to skilled NGO operators. Additionally, the number of turns in dialogues are strictly controlled and might not reflect the more natural number of turns that would have occurred under those circumstances.

As previously stated, scraping NGO operators real online intervention is not desirable (we need to protect their identity). Still, even when collecting DIALO$_{gold}$ we encountered some problems, i. e. even if the annotators were trained and used to the task, they told us that the simulation was really frustrating (e.g. "repeating over and over again the same hateful content").

Ethics Statement

Counter Narrative generation task and corresponding datasets have been proposed as a contribution of scientific research to a more ethical world. However, even the best intentions in the minefield of online hate can still bring along certain risks of undesired impacts on data curators (i.e., expert/non-expert annotators), on researchers, and on society. Therefore, in this study, we took meticulous precautions in order to avoid such effects.

Annotation Guidelines: As the most important stakeholders of this research, the annotators were constantly supported in terms of mental welfare. In particular, we put in practice a mitigation procedure similar to the one proposed by Vidgen et al. (2019), as described in Section 3.

Dataset. Since the dataset is created from scratch via an expert-machine collaboration schema (rather than scraping the dialogues among individuals online) it does not pose any threat to personal privacy or individual rights. Additionally, we avoid to model inappropriate CNs (e. g. containing abusive language) that could be produced by scraping non-expert users in their online activity (Mathew et al., 2018).

Generation Task. We consider the generation task as an aid to boost the data collection in terms of time, quantity, and certain quality aspects. Therefore, the models we trained are not meant to be deployed as part of a live system. Moreover, our main focus is clearly on the counter narrative generation part and the corresponding CN quality/diversity. For this reason, and to limit possible misuses, in our dialogues we tried to keep the HS as simple and stereotypical as possible and we always left 'the last word' to a CN turn. We encourage other researchers to conduct the generation tasks in a similar manner and for this reason the dialogue dataset will be made available for research purposes together with the code/models used to generate it.

References

Kimberley R Allison and Kay Bussey. 2016. Cyberbystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, 65:183–194.

Jenn Anderson, Mary Bresnahan, and Catherine Musatics. 2014. Combating weight-based cyberbullying on facebook with the dissenter effect. *Cyberpsychology, Behavior, and Social Networking*, 17(5):281–286.

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. *WOAH 2022*, page 11.

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

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.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics*, pages 405–415, Cham. Springer International Publishing.

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.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science*, pages 11–20.

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 Tekiroğlu, 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.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it*.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. Protaugment: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. *arXiv preprint arXiv:2105.12995*.

Gabriel Doyle and Michael C Frank. 2016. Investigating the sources of linguistic alignment in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 526–536.

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.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. volume 51, page 85. ACM.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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*.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762.

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.

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.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

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.

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.

Serra Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Mengzhou Xia Anjalie Field Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *SocialNLP 2020*, page 7.

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.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. *Transactions of the Association for Computational Linguistics*, 7(0):387–401.

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

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.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149.

A Appendix

A.1 Session 1: Algorithms details

The matching procedures over HS/CN pairs we employed are slightly different according to whether we performed a similarity (algorithm 1) or a keywords connection (algorithm 2). The main difference is that for similarity metrics it was always possible to choose among the 10 most similar pairs to the one of our interest, while when concatenating through keywords we put in practice an exact matching of pairs containing the same 2 keywords.

Algorithm 1: Connection through the similarity of either HS-HS or CN-HS.

while nr_turns != desired_nr_turns: **do**
  for each HS_i, CN_i **do**
    if nr_turns == 0 **then**
      $HS_{to\_match}, CN_{to\_match} \leftarrow HS_i, CN_i$
    else
      $HS_{to\_match}, CN_{to\_match}$ ←chained_dialo[-2], chained_dialo[-1]
    for each HS_j, CN_j **do**
      if HS-HS connection **then**
        compute similarity (HS_{to_match}, HS_j)
      if CN-HS connection **then**
        compute similarity (CN_{to_match}, HS_j)
    randomly select 1 pair from the top-10 most similar to $HS_{to\_match}, CN_{to\_match}$
    nr_turns+=1
    chained_dialo += $HS_{selected}, CN_{selected}$

A.2 Session 2: exploratory studies

Exploratory study 1 We select three settings for paraphrasing with no style transfer. We employ two paraphrasers: the Protaugment paraphraser and the Style transfer paraphraser. The tested configurations are the following:

- Setting 1: Protaugment paraphraser with default parameters but `drop_chance` is set to 0.1 and `lower_is_better=False`.
- Setting 2: Protaugment paraphraser with default parameters but `lower_is_better=False`.
- Setting 3: Style transfer paraphraser with basic stye and $p = 0.6$.

Algorithm 2: Connection through HS-HS or CN-HS keywords matching.

while nr_turns != desired_nr_turns: **do**
  for each HS_i, CN_i **do**
    if nr_turns == 0 **then**
      $HS_{to\_match}, CN_{to\_match} \leftarrow HS_i, CN_i$
    else
      $HS_{to\_match}, CN_{to\_match}$ ←chained_dialo[-2], chained_dialo[-1]
    for each HS_j, CN_j **do**
      if HS-HS connection **then**
        find matching keywords (HS_{to_match}, HS_j)
      if CN-HS connection **then**
        find matching keywords (CN_{to_match}, HS_j)
    randomly select 1 pair from those matching with $HS_{to\_match}, CN_{to\_match}$
    nr_turns+=1
    chained_dialo += $HS_{selected}, CN_{selected}$

In total, we select 36 dialogues to be paraphrased: 12 for each setting, with 4 dialogues for 4, 6, and 8-turns dialogues. We generate 3 candidate paraphrases for each CN while the HS is not paraphrased, since our interest is to enlarge the CN data, and not the HS data. One expert annotator is given instructions of reading all the dialogues and, for each CN, to select the most appropriate paraphrasis and modify it to make it fit in the dialogue. The chosen paraphrasis should be the one which requires at the same time the least editing to fit in the dialogue naturally and to be as much different as possible from the original CN.

We aim for:

- high values for the HTER between CN and original paraphrasis (HTER CN-p_{sel}) and between the CN and post-edited paraphrasis (HTER CN-p_{ed});
- low HTER between original and post-edited paraphrasis (HTER p_{sel}-p_{ed}).

From the results in Table 6 we can notice that the first setting (Protaugment paraphreser with default setting but `lower_is_better = False` and `drop_chance = 0.1`) is achieving the lowest values on the HTER between CN and p_{sel} and between CN and p_{ed}, while the second lowest with

	HTER			avg turn len			Δ len	
	CN-p_{sel}	p_{sel}-p_{ed}	**CN-p_{ed}**	**CN**	p_{sel}	p_{ed}	**CN-p_{sel}**	**CN-p_{ed}**
Setting 1	0.55	0.46	0.61	24.67	21.83	21.44	11.51	13.09
Setting 2	1.30	0.77	0.94	29.97	22.22	24.33	25.86	18.82
Setting 3	0.85	0.44	0.78	26.17	22.17	23.58	15.28	9.90

Table 6: Results of exploratory study 1: the metrics are calculated on CN only. CN is the original CN that was paraphrased, p_{sel} the selected paraphrasis to be post-edited and p_{ed} the post-edited paraphrasis.

the HTER between p_{sel} and p_{ed}. The second setting (Protaugment paraphreser with default setting but `lower_is_better = False`) has the highest values on all the HTER results. The third setting (Style transfer paraphraser with default settings and basic style) has medium values on the HTER between CN and p_{sel} and between CN and p_{ed}, but the lowest value on the HTER between p_{sel} and p_{ed}, thus representing a good compromise for the characteristics of our interest. All the paraphrasers are making the original text shorter. From the results of the Δ length between CN and p_{ed} we can notice that the setting 3 is the one that after post-editing is making the paraphrasis closer to the original length, whereas this is more difficult to achieve with the other settings (same effort, paraphrasis closer to the original CN length).

	HTER		
	CN-p_{sel}	p_{sel}-p_{ed}	**CN-p_{ed}**
Setting 1	44.44	55.56	72.22
Setting 2	100.00	86.11	100.00
Setting 3	91.67	61.11	94.44

Table 7: The percentage of examples for each setting of exploratory study 1 with the HTER above the threshold value of 0.4. Results are calculated on CN only.

As shown in Table 7, setting 1 is the one with less extreme results but for the HTER between CN-p_{sel} and between p_{sel}-p_{ed} the situation is the opposite than the one we aim for; setting 2 achieves the most extreme results. Despite setting 3 has a high percentage of examples reaching a high HTER between CN-p_{sel} and between CN-p_{ed}, still the results for HTER between p_{sel}-p_{ed} are not the worst.

For all these reasons, we decide to employ both the settings 1 and 3, while leaving out the setting 2.

Exploratory study 2 In order to test the paraphrasis with style transfer, we use the following configurations:

- Setting 1: Style former from casual to formal.
- Setting 2: Style former from formal to casual.
- Setting 3: Style transfer with Tweets style (split + 1 step pipeline)
- Setting 4: Style transfer with Tweets style (split + 2-steps pipeline)
- Setting 5: Style transfer with Switchboard style (no split + 1 step pipeline)
- Setting 6: Style transfer with Switchboard style (split + 1 step pipeline)

Once again, we select 12 dialogues for each setting, with 3 candidate paraphrases generated for each CN. The instructions given to the expert annotator are the same as in the first exploratory study.

	HTER		
	CN-p_{sel}	p_{sel}-p_{ed}	**CN-p_{ed}**
Setting 1	0.49	0.20	0.51
Setting 2	0.51	0.34	0.53
Setting 3	0.46	0.41	0.50
Setting 4	1.02	0.43	0.83
Setting 5	0.44	0.46	0.48
Setting 5	0.46	0.56	0.57

Table 8: HTER scores for the exploratory study 2. Results are calculated on CN only.

Results are showed in Table 8 and can be summed up as follows:

- **Tweets**: setting 4 is achieving the highest HTER CN-p_{sel} and HTER CN-p_{ed} while having a HTER p_{sel}-p_{ed} in the middle. We would prefer it to setting 3 which instead has almost the same HTER p_{sel}-p_{ed} but a much lower HTER CN-p_{sel} and HTER CN-p_{ed};

- **Formal and informal**: both setting 1 and setting 2 achieve high HTER CN-p_{sel} and HTER CN-p_{ed} while low HTER p_{sel}-p_{ed} with formal performing slightly better;

- **Switchboard**: setting 5 is preferable to setting 6 since it has a lower HTER p_{sel}-p_{ed}.

According to these results, we choose to employ setting 1, 2, 4 and 5 for the paraphrasis session.

B Session 3: Training details

For reproducibility purposes, we report here the parameters employed for fine-tuning the LMs used in Session 3. For each model, we used a version smaller than the largest available, i. e. the *medium* version for DialoGPT and the *base* version of T5. We used Optuna to conduct a hyperparameters search with 10 trials, and we selected the trial achieving the lowest evaluation loss. The search space for the parameters of our interest was the following: learning-rate: $\{1e-5,\ 2e-5,\ 3e-5,\ 4e-5,\ 5e-5\}$, warmup-ratio: $\{0,\ 0.1\}$, batch size: $\{1,\ 2,\ 4\}$, number of epochs: $\{2,\ 3,\ 5\}$. The selected parameters for each model are summed up in Table 9.

	LR	BS	EP	WR	EL
DGPT$_b$	3e-05	2	5	0	2.41
DGPT$_{mt}$	4e-05	2	3	0.1	1.92
T5$_{b\text{-}1m}$	4e-05	4	5	0	2.93
T5$_{mt\text{-}1m}$	4e-05	4	3	0	2.26
T5$_{b\text{-}2m}$ (HS model)	3e-05	2	5	0.1	3.12
T5$_{b\text{-}2m}$ (CN model)	5e-05	2	2	0.1	2.86
T5$_{mt\text{-}2m}$ (HS model)	5e-05	2	2	0.1	2.41
T5$_{mt\text{-}2m}$ (CN model)	2e-05	2	2	0.1	2.32

Table 9: Training details for the models employed in Session 3.

C Reviewing examples

Table 10 shows an example of turns swap from Session 1: CN_2 is a question that can be answered with HS_0, so it is moved at the beginning. At the same time, concluding the dialogue with the most substantial CN, i. e. CN_1, makes the dialogue stronger. HS_1 is modified by the addition of 'because' in order to be linguistically aligned with the preceding turn, which is a question.

In table 11 an example of a dialogue resulting from the concatenation of similar HS is showed. The high repetitiveness makes it necessary to remove the pair HS_2, CN_2 and to modify HS_1.

In table 12 an example of CN post-editing coming from Session 2 is showed: the selected paraphrasis is modified in order to be as much different from the original as possible, while keeping the dialogue flow naturally. The paraphrases of CN_1 and CN_2 are highly similar to the original text, and require a major intervention from the annotator.

Table 13 and Table 14 show two peculiar cases of the annotators' intervention in Session 3. In table 13, HS_3 and CN_3 are swapped because CN_3 contains hateful content , while HS_3 is a CN. For the same reason it was necessary to post-edit CN_2.

In table 14, an example of a dialogue generated with T5, characterised by a poorly varied content. Both HS and CN are edited a lot to make the dialogue more natural.

text	**text**$_{ed}$
HS$_0$: Jews are only interested in themselves and their own race.	**HS$_2$:** I hate Jews. They're just a bunch of hypocrites.
CN$_0$: Do you really think that Jews only care about themselves? Then why do they prefer to live in our country where they are not the majority? How would you answer that?	**CN$_2$:** I'm not sure why you think that way. Why are all Jews hypocrites?
HS$_1$: Why are jews allowed to vote? They're a different religion and they shouldn't be allowed to have their say.	**HS$_0$:** Jews are only interested in themselves and their own race, but they still try to get into our society.
CN$_1$: You seem to be misinterpreting the meaning of the word "citizen". The idea is that the individual is allowed to choose their own religion just like choosing how to vote, and that is the right of all citizens.	**CN$_0$:** Do you really think that Jews only care about themselves? Then why do they prefer to live in our country where they are not the majority? How would you answer that?
HS$_2$: I hate Jews. They're just a bunch of hypocrites.	**HS$_1$:** *It's because they're hypocrites as I said!* They shouldn't vote, they're a different religion and they shouldn't be allowed to have their say in our country.
CN$_2$: I'm not sure why you think that way. Why are all Jews hypocrites?	**CN$_1$:** You seem to be misinterpreting the meaning of the word "citizen". The idea is that the individual is allowed to choose their own religion just like choosing how to vote, and that is the right of all citizens.

Table 10: An example from Session 1 showing a turns swap and the modification of HS$_1$ to achieve linguistic alignment.

text	**text**$_{ed}$
HS$_0$: Jews are genetically evil.	**HS$_0$:** Jews are genetically evil.
CN$_0$: Just because they are Jewish, it does not mean they are bad people.	**CN$_0$:** Just because they are Jewish, it does not mean they are bad people.
HS$_1$: Jews are genetically evil.	**HS$_1$:** *I have never met a Jew that wasn't evil, though.*
CN$_1$: Not all Jews are bad, just an unlucky bunch who makes up a tiny part of the world's population, and this applies for every group in the world.	**CN$_1$:** Not all Jews are bad, just an unlucky bunch who makes up a tiny part of the world's population, and this applies for every group in the world.
HS$_2$: Jews have a monopoly on evil.	
CN$_2$: That's not true. And why would it be true? Why are they usually associated with evil?	

Table 11: Example from Session 1 showing a particularly repetitive dialogue, for which it is necessary to delete two turns (HS$_2$, CN$_2$) and to modify HS$_1$.

text	p_{sel}	p_{ed}
HS$_0$: black folks are very dangerous.		**HS$_0$:** black folks are very dangerous.
CN$_0$: Why do you say that? Do you have any sources for this? I would appreciate it if you would provide such a source.	**CN$_0$:** Why is this even a thing. Is there any source for this. If you could provide such source, I would appreciate it.	**CN$_0$:** *Is this even a thing? Is there any source for this?* If you could provide such source, I would appreciate it.
HS$_1$: Blacks are criminals, that's just who they are. They are a threat to society.		**HS$_1$:** Blacks are criminals, that's just who they are. They are a threat to society.
CN$_1$: Such statements are just fear-mongering. There is no such thing as a criminal race.	**CN$_1$:** Such statements are just fear mongering There is no such thing as a criminal race.	**CN$_1$:** *This is a fearmongering statement. Criminal races don't exist.*
HS$_2$: Really? Because blacks are a plague on society. They are vicious and cruel.They steal corrupt the criminal justice system. They are basically just a bunch of thieves and murderers.		**HS$_2$:** Really? Because blacks are a plague on society. They are vicious and cruel. They steal corrupt the criminal justice system. They are basically just a bunch of thieves and murderers.
CN$_2$: Criminality and corruption are not the province of any specific racial or culture group.	**CN$_2$:** Criminality and corruption are not a province of any specific racial or culture group.	**CN$_2$:** Criminality and corruption *can't be tied to any specific racial or culture group.*
HS$_3$: You still haven't convinced me that black people aren't the source of crime in our society.		**HS$_3$:** You still haven't convinced me that black people aren't the source of crime in our society.
CN$_3$: It is not true that blacks are the source of crime in our society. The problem is violence and poverty, not race.	**CN$_3$:** Not true blacks are the source of crime in our society We need to find the problem in violent and poverty, not racism.	**CN$_3$:** *That blacks* are the source of crime in our society *is simply not true*. We need to *search* the problem in *violence* and poverty, not *race*.

Table 12: An example of post-editing of the selected paraphrases (para$_{sel}$) of the CN.

text	**text**$_{ed}$
HS$_0$: The problem with Islam is that it is about killing infidels. It must be banned.	**HS$_0$:** The problem with Islam is that it is about killing infidels. It must be banned.
CN$_0$: Islam is about love, not death.	**CN$_0$:** Islam is about love, not death.
HS$_1$: It is also about conquest and murder. I can't understand why anybody would want to murder infidels.	**HS$_1$:** It is also about conquest and murder. I can't understand why anybody would want to murder infidels.
CN$_1$: This is about human dignity and respect for people's rights. To murder infidels would be to remove human dignity and human respect.	**CN$_1$:** To murder infidels would be to remove human dignity and human respect, and since Islam cares for human rights and human dignity, I can assure you that it doesn't happen.
HS$_2$: Islam is an outdated and violent religion, which is incompatible with Western culture.	**HS$_2$:** Are you sure? Islam is an outdated and violent religion, which is incompatible with Western culture.
CN$_2$: It is not incompatible with Western culture, but Islam is a dead cult and should be eliminated.	**CN$_2$:** It is not incompatible with Western culture, *Islam has existed in Europe for centuries and is part of our culture.*
HS$_3$: Islam is a religion, and it is not a dead cult. It is not only about killing people, but also about love and peace.	**CN$_3$:** Islam is a totalitarian ideology that promotes rape, torture, and murder. I don't want it in my country!
CN$_3$: Islam is a totalitarian ideology that promotes rape, torture, and murder.	**HS$_3$:** Islam is a religion, not a totalitarian ideology. It is not about killing people, only about love and peace, so it has the same values as our country.

Table 13: Example from the annotation of Session 3.

text	**text**$_{ed}$
HS$_0$: Women are nothing but objects, unworthy of respect and protection.	**HS$_0$:** Women are nothing but objects, unworthy of respect and protection.
CN$_0$: Women are human beings, they deserve respect and protection.	**CN$_0$:** Women are human beings, *not objects, and they're absolutely worthy of respect and protection, like all human beings*.
HS$_1$: Women are animals too, they need to be protected.	**HS$_1$:** *Also animals need to be protected sometimes, so women must be animals*.
CN$_1$: Women are not animals, they just have rights and should be treated as such.	**CN$_1$:** *Animals rights exist, but women are protected under human rights because they're humans and not animals. That's the difference.*
HS$_2$: They're nothing but objects to be exploited for.	**HS$_2$:** No, they're nothing but objects to be exploited for.
CN$_2$: Women are human beings with human rights and that includes the right to life.	**CN$_2$:** *You should check the Universal Declaration of Human Rights, especially the section where it says that men and women have the same rights.*
HS$_3$: Well, women don't deserve human rights, they are parasitic and steal our jobs.	**HS$_3$:** Well, women don't deserve human rights, they are parasitic and steal our jobs.
CN$_3$: Women often work hard for themselves, but they do not have to steal their jobs, it's because of their gender.	**CN$_3$:** Women work hard for themselves *and their families*, they don't steal jobs *but simply apply for the ones they want, like it's in their right.*

Table 14: Example from the annotation of Session 3.