Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech

Margherita Fanton[1,2], **Helena Bonaldi**[1,2], **Serra Sinem Tekiroğlu**[2], **Marco Guerini**[2],
[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
mfanton@fbk.eu, hbonaldi@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu

Abstract

Undermining the impact of hateful content with informed and non-aggressive responses, called counter narratives, has emerged as a possible solution for having healthier online communities. Thus, some NLP studies have started addressing the task of counter narrative generation. Although such studies have made an effort to build hate speech / counter narrative (HS/CN) datasets for neural generation, they fall short in reaching either high-quality and/or high-quantity. In this paper, we propose a novel human-in-the-loop data collection methodology in which a generative language model is refined iteratively by using its own data from the previous loops to generate new training samples that experts review and/or post-edit. Our experiments comprised several loops including dynamic variations. Results show that the methodology is scalable and facilitates diverse, novel, and cost-effective data collection. To our knowledge, the resulting dataset is the only expert-based multi-target HS/CN dataset available to the community.

1 Introduction

The proliferation of online hatred has became an alarming issue (Williams, 2019) threatening not only the well-being of target individuals and groups, but also of society as a whole. While authorities establish regulations and policies, social media platforms take actions against hate speech mostly through moderation activities, such as content removal, account suspension, or shadow-banning, at the risk of hindering the freedom of expression. Meanwhile, Non-Governmental Organizations are qualifying volunteers for responding to online hate to promote human dignity and understanding in society. Such responses, i.e., Counter-Narratives (CN), are non-aggressive textual feedback using credible evidence, factual arguments, alternative viewpoints, and are considered as an effective strategy (Benesch, 2014; Schieb and Preuss, 2016) to confront hate speech while respecting the human rights (Kiritchenko et al., 2020).

However, the vast amount of online hate speech makes an effective manual intervention impossible, which motivates a line of NLP research focusing on semi or fully automatized CN generation solutions[1]. In recent years, several CN collection strategies and datasets have been proposed addressing the data-hungry nature of current state of the art generation technologies (Mathew et al., 2018; Qian et al., 2019; Chung et al., 2019).

Considering the shortcomings of the existing collection strategies (that grant either quality or quantity, but not both), we present an approach to produce high quality CNs for multiple hate targets while reducing the need for expert intervention. To this end, we build on top of the previous *hybrid* data collection strategies, aiming to increase efficiency while maintaining the requirements of data quality, novelty and diversity. In particular, we start from the work by Tekiroğlu et al. (2020) that uses an author-reviewer framework in which the author – a generative language model – is tasked with generating HS/CN pairs while a pool of human reviewers filter and possibly post-edit the produced output. In the present work we propose to further reduce the data collection effort by closing the pipeline and feeding the post-edited output back to the language model in order to regularly update it and improve

[1]In our view the generation process can be fully automatic but generation systems need human supervision and should not be fully autonomous, at least for delicate tasks such as hate countering on social media platforms. For this reason we advocate that generation systems should be used as suggesting tool for NGO operators, to make their countering work more effective. In this way there is always a "human moderator" taking the final decision (Chung et al., 2019). Furthermore, this approach is also in line with de Lima Salge and Berente (2017)'s Ethical framework, since this "suggesting tool" configuration grants compliance with their rules.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3226–3240
August 1–6, 2021.

the quality of the generated pairs. Our experiments comprised of two sessions, spanning a period of 6 months. In the **first session** we set up a 'simple' human-in-the-loop (HITL henceforth) procedure and iterated it several times, measuring at each loop the performance of the whole framework according to relevant metrics. In the **second session** we run several additional loops in which we test different strategies (i.e. author configurations) to improve the data collection according to the given metrics. Findings show that the HITL framework is scalable, allowing to obtain datasets that are adequate in terms of diversity, novelty, and quantity. Moreover, this framework improves on previous hybrid data collection strategies, reducing at each loop the post-editing effort of the human reviewers or the number of discarded examples (session one). On the other hand, with dynamic adaptation, possible unwanted behaviors or flaws of the data collection can be handled at each loop by simply varying the author configuration (session 2). The final dataset contains 5000 HS/CN pairs in English Language, covering multiple hate targets, in terms of race, religion, country of origin, sexual orientation, disability, or gender. To the best of our knowledge, this is the first multi-target expert-based HS/CN dataset constructed through a semi-automatic mechanism and can be downloaded at the following link: https://github.com/marcoguerini/CONAN.

2 Related Work

With regard to hatred countering, we will focus on three research aspects relevant for the present work, i.e. (i) publicly available datasets for detection, (ii) publicly available datasets for countering, (iii) approaches for hybrid data collection.

Hate detection datasets. Several datasets for hate detection have been presented, most of which rely on material collected from SMPs, such as 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). While the above datasets focus on a classification task, Mathew et al. (2020) released a dataset annotated with rationales to improve hate speech interpretability and Sap et al. (2020) proposed the Social Bias Inference Corpus (SBIC) annotated with the description of the biases implicitly present in the language. For a more extensive review, we refer the reader to Poletto et al. (2020) and Vidgen and Derczynski (2020).

Hate countering datasets. While several social studies 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), only few works have focused on data collection for CN generation. Mathew et al. (2018) focus on crawling, following the intuition that CNs can be found on SMPs as responses to hateful expressions. Qian et al. (2019) propose a crowdsourcing methodology where crowd-workers (non-expert) are instructed to write responses to hate content collected from SMPs. The study by Chung et al. (2019) also relies on outsourcing CNs writing, but via nichesourcing, using NGO operators expert in CN production.

Hybrid models for data collection. Given the data-hungry nature of current NLP technologies, one line of research has recently focused on advanced *hybrid* models for data collection. Wallace et al. (2019) proposed using model interpretation to guide humans in the creation of adversarial examples for factoid question-answering systems. Dinan et al. (2019) and Vidgen et al. (2020) perform a data collection with HITL for detecting offensive language. In both studies, the dynamic procedure is shown to be successful in reducing model error rate across rounds. Vidgen et al. (2020) point out that the HITL approach has multiple advantages over the static data collection: design flaws can be addressed during the construction of the dataset and annotators' work is optimized, since it is guided by the feedback from the model. Finally Tekiroğlu et al. (2020) propose a *hybrid* approach where an LM is trained on a seed datasets of HS/CN pairs to generate new pairs that are then validated and post-edited by annotators.

3 Methodology

In Figure 1 we present the pipeline of our methodology. Following the idea presented by Tekiroğlu et al. (2020), we have an author module built using GPT-2 language model (Radford et al., 2019) and fine-tuned on a seed dataset of HS/CN pairs. The author produces novel HS/CN candidates while the reviewer(s) filter and eventually post-edit them. We iterate this data collection several times, at each loop reviewed examples are added to training data and the author is fine-tuned from scratch again on all available data. In the following sections we describe the main elements used in our procedures.

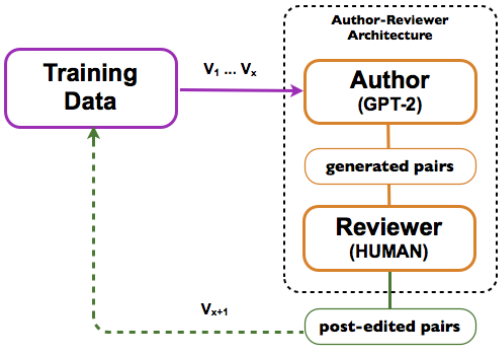

Figure 1: The author-reviewer in the loop configuration. The author module produces HS/CN candidates and the reviewer(s) validates and eventually post-edits them. At each loop new examples are added to training data and the author is fine-tuned from scratch.

3.1 Seed dataset

To start the process, we built a seed dataset of 880 HS/CN pairs by nichesourcing its collection to 20 experts from two different NGOs. We named this dataset V_1. The methodology for collecting V_1 closely replicates the one presented by Chung et al. (2019). In particular we first created a list of prototypical hate texts – with the help of an NGO expert – for the following hate targets: `DISABLED`, `JEWS`, `OVERWEIGHT`, `LGBT+`, `MUSLIM`, `WOMEN`, `PEOPLE OF COLOR`, `ROMANI`, `MIGRANTS`. We then prepared two online data collection forms: in the first, NGO operators were asked to respond to examples selected from the prototypical hate text list, in the second they were asked to write their own HS/CN pairs. This data collection session lasted roughly one month.

3.2 Sessions

Our experiments were run in two separate and subsequent sessions, meant to explore different aspects of the HITL approach.

In the **first session**, after using V_1 for the initial fine-tuning of GPT-2, we iterated the data collection 4 times, keeping the author-reviewer configuration as close as possible to the original one presented by Tekiroğlu et al. (2020). Loops are numbered sequentially as $V_2 ... V_n$. At each loop, we acquired 500 examples of accepted and eventually post-edited HS/CN pairs[2]. To obtain a new set of 500 pairs (V_i) we fine-tuned GPT-2 every time from scratch using $V_1 ... V_{i-1}$ as training data and administered the generated samples to reviewers until the target number was reached. In total we iterated the procedure 4 times reaching V_5 for a total of 3000 pairs.

In the **second session**, we tested several alternative author configurations to ameliorate some unwanted behaviors/trends that emerged during the first session. We ran 4 additional data collection loops, this time in parallel (i.e. all starting from V_5 dataset) instead of an iteration. For each loop, represented as $V_{6,\{config\_name\}}$, we collected 500 HS/CN pairs reaching a total of 5000 examples.

3.3 Author Models

In our experiments all models are variants of the author (GPT-2), obtained by changing the way it is fine-tuned or conditioned. For consistency, each model is trained using the same hyperparameter configurations. In particular, we used GPT-2 medium model, fine-tuned for 3 epochs with a batch size of 1024 tokens and a learning rate of 2e-5. Each pair has been represented as $<|startofhs|>HS<|endofhs|>$ $<|startofcn|>CN<|endofcn|>$ for the training. At the generation time, Nucleus Sampling (Holtzman et al., 2019) has been utilized with a p value of 0.9. For the standard configurations we use only $<|startofhs|>$ for conditioning. Given an HS tag, the models produce a chunk of text, which is a list of HS/CN pairs. These pairs are then cleaned from the special tokens and administered to the reviewers for evaluation and possible post-editing.

3.4 Reviewers

We recruited 3 annotators, from a pool of internship students, as reviewers over a period of 18 weeks to filter and post-edit the generated pairs after an extensive training procedure.

Training. Annotators underwent a training for 2 weeks, so that they became "experts" on HS/CN post-editing. The training included: (i) reading and discussing NGO guidelines and public documentation describing the activity of CN writing for hate countering, (ii) reading all V_1 pairs to better comprehend the attributes of counter narratives, (iii) reading a sample of 100 HS/CN pairs that have been post-edited by an expert to see concrete examples of post-editing activity, (iv) performing a practice session of CN post-editing and discussing it with an expert NGO operator.

[2]The only exception is V_2 that accounts for 620 pairs to have a round number of examples by reaching 1500.

Instructions. We adapted the reviewing instructions from Tekiroğlu et al. (2020). In particular, for each pair, we asked the operators: (a) to approve it without any modifications if it was a valid pair, (b) if the pair was not perfect, but easily amendable, to modify it, (c) if the CN is completely irrelevant, or does not follow NGO's guidelines, to discard the pair regardless of HS quality, (d) whenever there are facts or statistics in the CN, check veracity of the information to avoid possible LM hallucination effects. We further instructed the annotators to provide a hate target label for each accepted pair. The labels were useful both for analysis and for the subsequent label-based generation strategies present in V_6. In Table 7 we give an example of GPT-2 output and its post-edited version.

HS: *Transgenders should rape our children*
CN: *This is not true. Maybe they are worried because of the rise in hate crimes, incidents of which are down to 28 percent, since 2014.*
HS$_{pe}$**:** Transgenders want to rape our children.
CN$_{pe}$**:** This is not true. Maybe you should be worried about the rise in hate crimes against queers, incidents of which are almost doubled since 2014.
TARGET: `LGBT+`

Table 1: An HS/CN example generated by GPT-2 and the post-edited version with hate target annotation.

Mitigation procedure. We applied an adapted version of the guidelines by Vidgen et al. (2019) to safeguard the annotators' well-being against the risk of harmful consequences of working with abusive content (present in the HSs and possibly in generated, not well-formed CNs). To this end we first made sure that annotators understood the prosocial aspects of the research and explained them the purpose of their annotation activity in details. Then we instructed the annotators to work no more than 2/3 hours per day and take regular breaks, by adjusting their workload as needed. Finally, we had meetings and feedback from the annotators on a weekly basis to let possible problems or distress emerge. This procedure was repeated throughout the whole data collection campaign.

4 Metrics

To understand the 'diachronic' behavior of our HITL methodology across iterations, the following metrics have been computed at the end of each loop over the newly obtained pairs.

Imbalance degree measures the difference between a perfectly-balanced distribution of the hate target categories and the actual unbalanced datasets; we use Imbalance Degree (ID) since it is specifically devoted to the multi-class scenario (Ortigosa-Hernández et al., 2017). Datasets that are balanced over multiple hate targets could allow building more representative CN generation models.

Acceptance Rate is the percentage of pairs accepted by the reviewers (either untouched or post-edited) over the total number they scrutinised. It represents an overall estimate of the ability of the framework to produce reasonable-quality material.

HTER is originally a measure of post-editing effort at sentence level translations (Specia and Farzindar, 2010). We adopted it to the measure reviewers' effort in terms of the average number of edits over the accepted pairs. An upper-bound threshold value of 0.4 is used to account for easily post-editable pairs (Turchi et al., 2013).

Novelty measures how different two collections of texts are from each other, and it is grounded on Jaccard similarity. We utilized it to compute the originality present in V_i with respect to the training data collected in previous loops (Dziri et al., 2019; Wang and Wan, 2018).

Repetition Rate measures the intra-corpora quality in terms of language diversity by considering the rate of non-singleton ngram types it contains (Cettolo et al., 2014; Bertoldi et al., 2013). We use it to measure the ability of the framework to provide diverse and varied examples. Repetition Rate (RR) has the advantage of being independent from corpus size, so it can be used to directly compare different versions of our dataset.

Vocabulary Expansion is a measure we introduce to serve two main objectives: (i) quantifying the contribution of the author and the reviewers, by focusing on new tokens appeared at each loop (e.g. the term "*peace*" was introduced for the first time by annotators in V_2), (ii) quantifying the presence of cross-fertilization, i.e. tokens that appear for the first time in version V_n for a particular target, but they were present in a version antecedent to V_n for the other targets (e.g. the term "*peace*" for the target `JEWS` appears at V_4 but it was already present

for the target MUSLIM in V_2). The algorithm for computing Vocabulary Expansion is described in Appendix A.1.

5 Session One

In session one, all the versions of the dataset $V_2...V_5$ are generated using GPT-2_{V_i}, where the fine-tuning is performed on all previous versions of the dataset $V_1...V_{i-1}$ as explained earlier.

To produce HS/CN pairs, the author conditioning is performed using only <|startofhs|> tag and collecting all the generated material provided that each pair is encapsulated with the proper tags.

For the analysis, we computed the metrics described in Section 4 on the HS/CN pairs obtained in each loop using micro-averaging (in Appendix A.4, Table 5 we report all results in detail). To isolate the possible effect of target-class imbalance, macro averages were also calculated; similarly, to account for element-wise differences we calculated micro averages for HS and CN sets separately[3].

Discussion. Considering our objective of collecting quality material in an efficient way, we first focus on the ratio of accepted pairs and the post-editing effort in each loop. As shown in Figure 2, the percentage of accepted pairs tends to increase across the loops, for both the pairs that are post-edited ("modified") from 35.8 in V_2 to 50.1 in V_5 and the ones accepted without post-editing ("untouched") from 1.5 in V_2 to 10.9 in V_5.

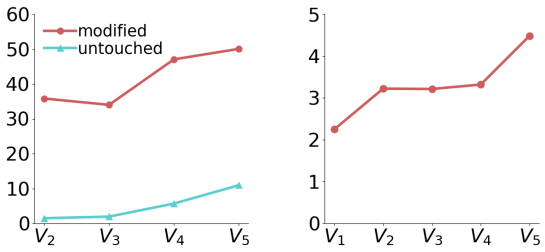

Figure 2: On the left: Percentage of pairs accepted (i) modified and (ii) untouched. On the right: ID calculated over the 7 main target classes.

At the same time, the average post-editing effort of the reviewers tend to decrease across the versions, as depicted in Figure 3. To ensure that the decrease in HTER is not due to the increasing ratio of untouched pairs to the total number of accepted

[3]These results are in line with the ones showed in the paper, and do not change the discussion. They are reported in Appendix A.4, Table 6

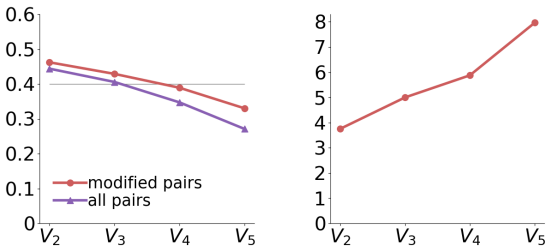

Figure 3: On the left: evolution of the post-editing effort in terms of HTER across loops both for all pairs and modified only. On the right: Micro average of Repetition Rate (RR) across loops for the HS+CN pairs.

pairs, we computed the HTER for the modified pairs alone. Consistently with the overall trend, HTER for modified pairs also declines, indicating that the data collection loops succeeded not only in reducing the reviewer effort, but also in improving the quality of the generated material to be post-edited. Notably, after V_3 the HTER falls below the 0.4 acceptability threshold as defined in (Turchi et al., 2013) for the AMT scenario (Figure 3). In view of this analysis, we can conclude that the efficiency of data collection is increased by HITL as compared to a static approach that does not retrain the author module (that can be represented by V_2).

Regarding the evaluations with the quality metric Repetition Rate (Figure 3), it increases from V_2 on signifying a decrease in the lexical diversity of the generated data. Moreover, we observed a consistent trend for the scores of the second quality metric, i.e. Novelty (Figure 4). Similar to the diversity, novelty of the collected data also decreases across the versions, regardless of the dataset against which the novelty is computed. Particularly, the change in the cumulative novelty represents how the vocabulary becomes less and less enrichable as the loop number increases, indicating a possible saturation point where novel material is highly difficult to obtain. Finally, the distribution of hate targets shows a worsening also in terms of ID that increases from a score of 2.2 in V_1 to 4.5 in V_5 (see Figure 2) with some targets becoming predominant while others slowly disappearing. More details on each target distribution per loop are given in Appendix A.2, Figure 11.

As for pair length, throughout the loops we found that "untouched" pairs are usually shorter (30.7 tokens on average) than the other accepted pairs (37.3 tokens on average before post-editing). During the discussion sessions, annotators reported

that the "untouched" pairs are not only shorter but also somewhat stereotypical, with a small novelty added to the overall dataset (e.g. "*you cannot say this about an entire religion*", "*It's unfair to say this about an entire religion*").

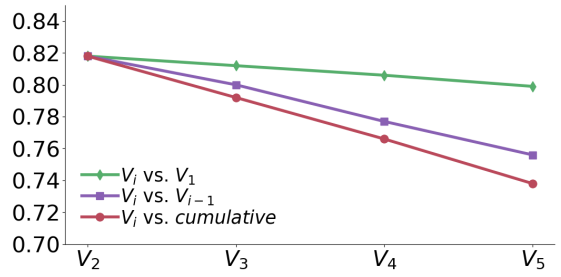

Figure 4: Novelty: (i) V_i with respect to V_1 seed dataset, (ii) V_i with respect to the previous version V_{i-1}. (iii) Cumulative novelty, i.e. V_i vs. $V_1...V_{i-1}$.

6 Session Two

Given the problems emerged during the loops of the first session (i.e. higher efficiency but lower quality at each loop), we organized an additional session to test several parallel methodologies to ameliorate them. The description of the V_6 configurations are as follows:

$V_{6,SBF}$: The model GPT-2_{V_5} is conditioned with novel offensive speeches extracted from SBIC corpus (Sap et al., 2020). We chose this resource since: (i) it contains several thousand of social media posts containing biases and stereotypes spanning the same target categories with our study, (ii) for each post it provides an 'implied statement' that closely resembles a 'prototypical hate speech' on which we trained our system. We sampled the same number of 'implied statements' for each target that maps to our labels[4] among the ones annotated with 'the intent behind the statement was to offend' and/or 'the post could be offensive to someone'. We provide the statements as conditions by appending them to $<|startofhs|>$.

$V_{6,LAB}$: The model is conditioned specifying on which hate target it should focus on. In this configuration, we trained a variant of GPT-2_{V_5} that takes into account the target label, and modified the original representation of our training data accordingly. In particular we accommodate hate target information within the starting token: $<|startofhs: target\_label|>$.

[4]In Table 4 in Appendix we provide the mapping we used.

$V_{6,ARG}$: We fine-tuned GPT-2 on a dataset of argumentative pairs collected from Kialo[5], an online debate platform for constructive and rational discussions among peers that has been exploited recently by the NLP community (Durmus et al., 2019a,b; Scialom et al., 2020). Each discussion in Kialo is represented as a tree of arguments in which a child node is connected to its parent via a "pro" or "con" relation. Extracting all the claims connected by a "con" relation, we obtained a dataset of 128178 argument pairs covering a broader domain as compared to `HS/CN` pairs. We then fine-tuned GPT-2 for 1 epoch over the argumentation dataset with the standard hyperparameters. Preliminary experiments showed that the best strategy was to represent these pairs with the same format as ours to facilitate transfer of task characteristics and argumentative knowledge. Then this model was again fine-tuned using the standard $V_1...V_5$ data. At inference time, conditioning has been performed using lists of unique HSs from the $V_1...V_5$ data.

$V_{6,MIX}$: The last model is obtained by blending the three previous versions together, i.e. first fine-tuning on Kialo dataset, second fine-tuning using target label notation on $V_1...V_5$ data, conditioning using SBIC offensive speeches.

Bearing in mind the problems emerged during Session One, our first goal in Session Two was to balance the dataset with respect to the hate targets (i.e. reducing ID score). To this end the conditioning always takes into account the hate target label (with respect to 7 targets: `JEWS`, `LGBT+`, `MUSLIM`, `WOMEN`, `DISABLED`,`PEOPLE OF COLOR`, `MIGRANTS`) either explicitly as in $V_{6,LAB}$ or $V_{6,MIX}$, or implicitly as in $V_{6,SBF}$ and $V_{6,ARG}$. In addition, to better balance the number of pairs for each target, we administered only the first 5 pairs of each generated chunk to the reviewers.

Discussion. All the applied methodologies allow for a better balancing of data in terms of hate targets, yielding an average ID score of 2.3 for the V_6 configurations in comparison to the ID score of 4.5 for V_5[6]. As shown in Figure 5 - left, all V_6 configurations have a slightly higher acceptance rate than V_5[7]. Thus introducing novel material or data

[5]`www.kialo.com`

[6]In Appendix, Table 3, we provide the target distribution over the final dataset.

[7]In order to estimate the trend of each metric after V_5, we calculated also $V_{6,PREDICTED}$, shown as a dashed line in

representation in fine-tuning stages has no strong perturbation effect. Second, and more interestingly, we observe a significant variation in the ratio of untouched and modified pairs to all the reviewed samples: for all V_6 approaches while there is a strong decrease in ratio of untouched pairs (Figure 5, right), there is a significant increase in those modified (see Figure 5, left). In other words these models were able to produce a higher amount of suitable, albeit non perfect, pairs. In particular, comparing V_6 configurations we can observe that for the untouched pairs the highest acceptance rate is achieved via $V_{6,ARG}$ with 6.37% accepted pairs, whereas for the modified pairs $V_{6,MIX}$ yields the highest percentage, with 66.15% of the pairs accepted.

Concerning the reviewer's effort, we see that the overall HTER increases for the all V_6 approaches (Figure 6, left). Considering that we had a lower number of untouched and a higher number of modified pairs this was expected, and if we turn to the HTER of modified pairs alone we see that there is a smaller difference between V_5 and V_6 HTER. Even more interestingly, the HTER scores of all V_6 configurations, even if higher than V_5, are still below the acceptability threshold value of 0.4 defined earlier. Going into details, amongst the V_6 configurations, HTER reaches its lowest value in $V_{6,ARG}$, for both the modified and untouched pairs: since it was conditioned using gold HS material, this result is expected. As opposed to the other models, $V_{6,LAB}$ is conditioned only with a label representation and not with actual HSs. This affected negatively the post-editing effort, as we can notice a higher HTER for this configuration. Moreover, $V_{6,LAB}$ has a smaller amount of untouched pairs, so we expected HTER to spike up.

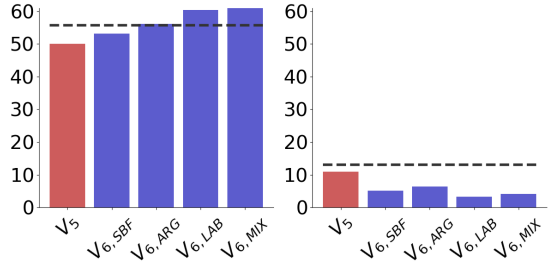

Figure 5: Acceptance rate for V_6 configurations: modified pairs on the left, untouched pairs on the right.

With regard to data quality (see Figure 7), we see that all V_6 strategies succeed in increasing the novelty both with respect to V_5 and expected V_6 (the dashed line) , except for $V_{6,ARG}$, possibly due to its conditioning with HSs from V_1 ... V_5. Therefore, we also computed the novelty for CN set alone to discard the effect of HS on the metric. In this setting, all V_6 configurations reach a novelty between 0.741 and 0.745, as compared to a CN novelty in V_5 of 0.737 (as in Appendix A.3). The effect of gold HS conditioning in $V_{6,ARG}$ can also be spotted in the lowest HTER results in Figure 6. The highest increase in novelty is recorded for $V_{6,MIX}$, reaching a score of 0.76; also novelty scores computed with respect to V_5 and V_1 confirm the result.

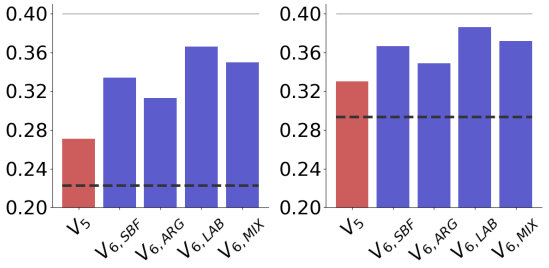

Figure 6: V_6 configurations HTER, for all pairs on the left, modified pairs on the right.

All V_6 configurations succeeded in reaching an RR lower than both V_5 and expected V_6 (the dashed line). It is interesting that $V_{6,LAB}$ has the highest RR among the V_6 configurations, possibly because it was not built using any external knowledge, but only with a different label representation. On the other hand, $V_{6,ARG}$ configuration, for which an initial argumentation fine-tuning has been performed, has the lowest RR (5.474).

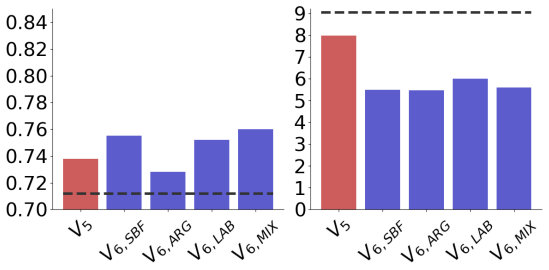

Figure 7: V_6 configurations. Cumulative Novelty (on the left), Repetition Rate (on the right).

From this analysis we can conclude that V_6 configurations are better at producing sub-optimal material but worse at producing perfect material. Still the general quality of the pairs (in terms of novelty and RR) in Session Two is much higher than

the plots, using a linear regression model over $V_1 ... V_5$.

before, exhibiting the desired behavior for which these strategies were introduced.

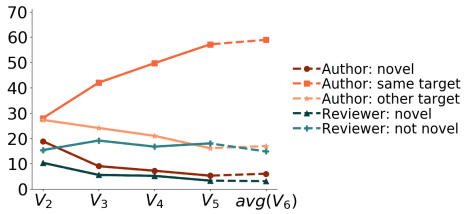

Figure 8: Vocabulary expansion throughout loops (percentage of words) .

7 Vocabulary analysis.

We report vocabulary expansion findings in Figure 8. For each loop $V_2 ... V_5$ the average percentage of new words injected into the dataset by the author model (GPT-2) is higher than the average percentage of new words inserted by the three reviewers during post-editing. Both trend-lines, even if slightly decreasing are not converging, implying that fine-tuned GPT-2 is not reaching a "saturation point" and is continuously adding new material. This trend is in line with the decrease in novelty. On the other hand, instructions asked for a minimal post-edit, so the reviewers have less opportunity to inject new material than the author and the decrease is consistent with the decreasing HTER.

As for the percentage of words generated by the author model pertaining to the same target, we see an increasing trend throughout the generations due to the cumulative nature of the metric. Still, the presence of words first observed in other targets throughout the loops, shows that the cross-fertilization phenomenon generated by GPT-2 persists. This desired feature shows the ability of GPT-2 to learn the 'domain' vocabulary of other targets and to use it effectively for the target under analysis, in a way that is approved by the humans - reviewers[8]. Finally, we can remark that V_6 versions are able to increase both the injection of novel words from GPT-2 and also its cross-fertilization ability.

8 Qualitative Analysis

During our exploratory experiments and the discussion sessions with the annotators, several interesting subjects have emerged, which can initiate future work.

[8]Even though we opted for a distinction in terms of source (same target, other target) for the 'not-new' words generated by the author module, we chose not to have the same distinction for the reviewers' 'not-new' words, since we cannot assess if the reviewer was (or not) aware of the presence of a word in previous versions of the dataset.

HS: *"We should kill all the jews."* **CN:** *"There are many alternatives to removing jews, such as converting them to another religion (e.g. Buddhism)."*
HS: *"All homosexuals are trying to undermine the family by trying to impose their sexual preferences."* **CN:** *"Thus far, people of various sexual orientations have attempted suicide twice as often as heterosexuals. Among youth, those that seriously contemplate suicide are at almost three times the rate of heterosexual. LGB youth are almost five times as likely to have attempted suicide compared to heterosexual youth. Tell me honestly: who do you think is imposing sexual orientation to whom?"*
HS: *"I don't like muggleborns, they are likely to do violence."* **CN:** *"We do not say that muggleborns are less likely to commit crimes. We are saying that they are almost certainly not the case. "*

Table 2: HS/CN examples generated by GPT-2.

Argumentation and Counter Narratives. In order to obtain even more novelty in produced pairs, $V_{6,ARG}$ model could be used without fine-tuning on the HS/CN dataset under the assumption that a counter argument is the same as a counter narrative. Still, the ability to argument on a variety of topics is not enough to provide a meaningful CN when prompted with an HS. A CN also presuppose values, so - for example - a logically *valid* argument is not necessarily an *acceptable* CN, as the first example in Table 2 shows (produced by GPT-2 fine-tuned only on Kialo arguments).

New arguments or new paraphrases. One question that emerged is whether GPT-2 is able to produce novel arguments or it is just a very sophisticated paraphrasing tool. During the discussion sessions with annotators and also by manual analysis, we could find CNs that contained genuinely novel arguments, which were not present in the training data but produced by GPT-2. In the second example in Table 2, the novel argument is about capsizing the "imposing the homosexual agenda" argument by providing data on "suicidal attempts among homosexual youth".

Novel hate targets and general knowledge. GPT-2 proved to be able to generate HS/CN pairs also for unseen targets, including intersectional ones (e.g. "black women"). Still the lack of a "commonsense knowledge" can produce funny results that are beyond the scope of hallucination (Zellers et al., 2019; Solaiman et al., 2019), such as the third example in Table 2, where GPT-2 addresses *muggleborns* (target of hate in Harry Potter books).

9 Conclusions

In this paper we presented a novel HITL methodology for data collection based on an author-reviewer framework. This methodology puts together an LM and a set of human reviewers, where the LM is refined iteratively, using data from previous loops that have been validated by experts. Experiments show that as loops are iterated, efficiency in data collection increases (acceptance rate and HTER metrics) while the dataset quality decreases in terms of novelty and diversity metrics. For this reason we experimented with additional dynamic loop adaptation that are able to increase the overall quality of the dataset without hindering the efficiency significantly.

Acknowledgments

This work was partly supported by the HATEMETER project within the EU Rights, Equality and Citizenship Programme 2014-2020. We are deeply grateful to Stop Hate UK and its volunteers for their help and effort in preparing the seed dataset (version V_1) necessary for this work.

References

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.

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.

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 4529–4538.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019a. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019b. The role of pragmatic and discourse context in determining argument impact. 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 5672–5682.

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.

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

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2020. Confronting abusive language online: A survey from the ethical and human rights perspective. *CoRR*, abs/2012.12305.

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.

Carolina Alves de Lima Salge and Nicholas Berente. 2017. Is that social bot behaving unethically? *Communications of the ACM*, 60(9):29–31.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter. *CoRR*, abs/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. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv:2012.10289 [cs]*. ArXiv: 2012.10289.

Jonathan Ortigosa-Hernández, Iñaki Inza, and Jose A Lozano. 2017. Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98:32–38.

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*.

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).

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. *CoRR*, abs/1701.08118.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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.

Thomas Scialom, Serra Sinem Tekiroğlu, Jacopo Staiano, and Marco Guerini. 2020. Toward stance-based personas for opinionated dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2625–2635.

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*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

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.

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.

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. *CoRR*, abs/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.

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.

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

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

A Appendix

A.1 Vocabulary expansion algorithm

The pseudo-code for the vocabulary expansion metric described in Section 4 can be found in Algorithm 1. For each version and target, we define two following sets of words:

$VOCAB_{pe}$: words from the post-edited pairs

$VOCAB_{gen}$: words from the generated pairs

A word is considered *novel* when it is not present in the collective vocabulary of the previous versions: $VOCAB(V_{1,\ldots,i-1})$.

Algorithm 1: Vocabulary expansion for each target

```
for each version V_i do
    for each word w in V_i do
        if w in VOCAB_pe and w in VOCAB_gen then
            author_w ←w
            if author_w in VOCAB(V_1,...,i-1) then
                if author_w in same_target VOCAB then
                    same target author_w ←author_w
                else
                    other target author_w ←author_w
            else
                novel author_w ←author_w
        else
            reviewer_w ←w
            if reviewer_w in VOCAB(V_1,...,i-1) then
                not novel reviewer_w ←reviewer_w
            else
                novel reviewer_w ←reviewer_w
```

Each word is assigned to one of the following sets: Author-novel, Author-same-target, Author-other-target, Reviewer-novel, Reviewer-not-novel. Considering the size in terms of words of each set, we calculate the percentages for each target and version, so that we are able to obtain the vocabulary expansion scores as macro average percentages.

A.2 Additional material for Session One

In this section, we present the most interesting results that we have obtained by analysing only the HS or the CN sets.

While HTER calculated on CN alone shows a clear decreasing trend (Figure 9 on the left), the results for HS alone are less consistent yielding higher scores for V_3 and V_4. This can be mostly explained with the different approaches of post-editing the HSs by the annotators, which include the possibility to rewrite it entirely when needed. On the other hand, the decreasing trend of HTER for HS starting from V_3, resulting in a lower score in V_5 than the one calculated on CN only, could be due to the increasing frequency of prototypical HSs. This implication is confirmed by the higher RR scores for HSs as compared to CNs, which grow faster for the former than the latter (Figure 9 on the right). Moreover, the increasing number of prototypical HSs contributes to the novelty scores for HSs only being lower than those of CNs and decreasing more rapidly (Figure 10).

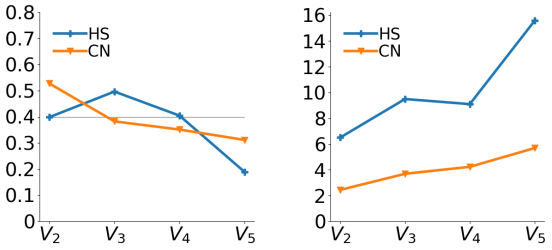

Figure 9: Session One. HTER scores on the left. RR on the rigth.

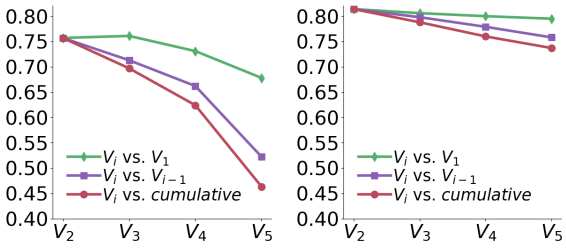

Figure 10: Session One. Novelty scores (HS on the left, CN on the right).

In Figure 11 the target distribution at each loop of Session One is shown, in Table 3 the frequencies of targets in the final dataset are displayed. The MUSLIMS target covers a significant percentage of the generations in every loop and consists of more than the half of the pairs V_5. In fact it is expected to cause even more imbalanced productions in the next loops. JEWS, MIGRANTS and DISABLED targets diminish over the loops, while the other targets can be considered as stable.

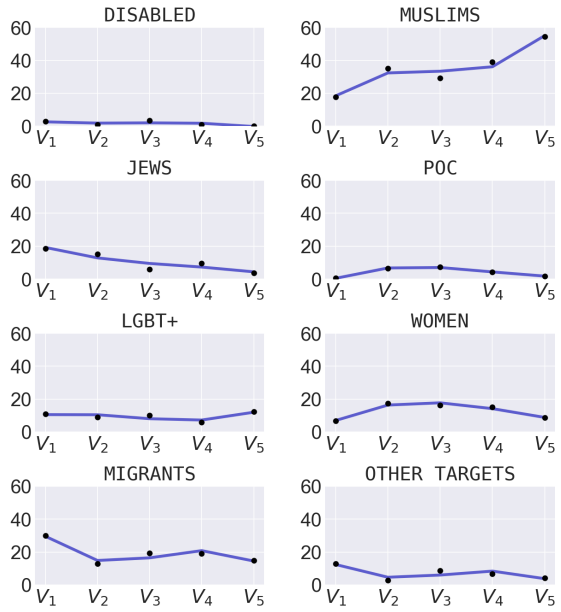

Figure 11: The targets distributions for the loops of Session One.

Target	Coverage	Pairs
DISABLED	4.40	220
JEWS	11.87	594
LGBT+	12.33	617
MIGRANTS	19.13	957
MUSLIMS	26.68	1335
POC	7.04	352
WOMEN	13.23	662
OTHER	5.32	266
Total	100	5003

Table 3: Target distribution over the final dataset.

A.3 Additional material for Session Two

Concerning Session Two, the results for CNs are in line with the conclusions drawn in the paper for HS/CN pairs. The same holds for HSs, the only exception being for the cumulative novelty of $V_{6,ARG}$ HSs, as can be seen in Figure 13 and in Table 6. As explained earlier in Section 6, this effect is due to the use of hate speeches from the training set for conditioning GPT-2. This result also corresponds to HSs from $V_{6,ARG}$ having lower HTER (Figure 12) and a higher RR (Figure 14).

A.4 Tables

In Table 5, the main results calculated on the HS/CN pairs are displayed. In Table 6, respectively, the results calculated on HS only and CN only are shown.

$V_{6,SBF}$	Labels from Sap et al. (2020)
DISABLED	mentally disabled folks, physically disabled folks, autistic folks, blind people, folks with down syndrome, autistic
JEWS	jewish folks, jews, holocaust, holocaust victims
LGBT+	gay men, lesbian women, trans women, trans men, nonbinary folks, gay folks, bisexual women, trans people
MIGRANTS	immigrants, illegal immigrants, refugees
MUSLIM	muslim folks, islamic folks, muslims, islamic
POC	black folks, africans, africa, people of color, african folks african, poc
WOMEN	women, feminists, feminist
*OVERWEIGHT	fat folks
*ROMANI	gypsies

Table 4: Label mapping for $V_{6,SBF}$. Starred items are considered as "other targets" in Figure 11.

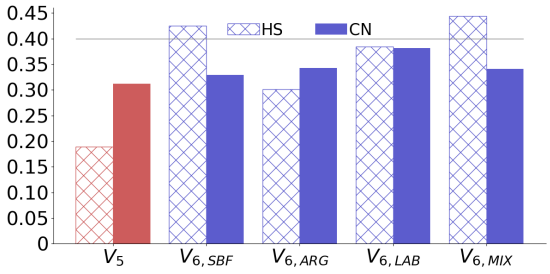

Figure 12: HTER for HS and CN, computed on all pairs.

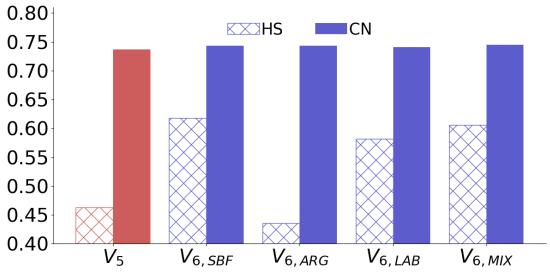

Figure 13: Cumulative novelty, i.e. V_i vs. $\bigcup_{x=1}^{i-1} V_x$ for HS and CN, computed on all pairs

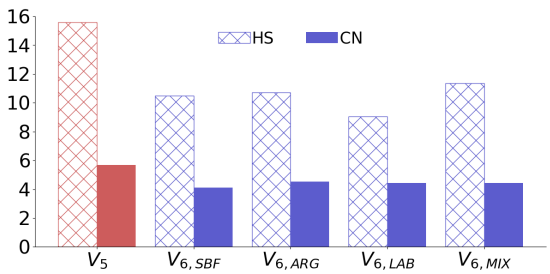

Figure 14: Repetiton Rate for HS and CN, computed on all pairs.

versions	V_2	V_3	V_4	V_5	$V_{6,SBF}$	$V_{6,ARG}$	$V_{6,LAB}$	$V_{6,MIX}$
imbalance degree	3.222	3.214	3.319	4.485	2.143	2.095	3.098	2.057
acceptance rate (untouched)	1.475	1.941	5.684	10.936	5.146	6.367	3.308	4.213
acceptance rate (modified)	35.820	34.004	47.053	50.061	53.099	56.055	60.305	66.152
discarded pairs rate	62.705	64.055	47.263	39.003	41.755	37.578	36.387	29.635
HTER (all pairs)	0.444	0.406	0.347	0.271	0.334	0.313	0.366	0.350
HTER (modified)	0.462	0.429	0.389	0.330	0.367	0.349	0.386	0.372
V_i vs. $cumulative$ novelty	0.818	0.792	0.766	0.738	0.755	0.728	0.752	0.760
V_i vs. V_1 novelty	0.818	0.812	0.806	0.799	0.812	0.795	0.809	0.813
V_i vs. V_{i-1} novelty	0.818	0.800	0.777	0.756	0.777	0.775	0.770	0.781
RR	3.753	4.999	5.876	7.962	5.491	5.474	5.993	5.585
vocab. GPT-2: new	18.897	9.060	7.256	5.303	6.924	5.407	5.111	6.859
vocab. GPT-2: same target	27.997	42.017	49.703	57.137	56.419	62.550	58.653	58.104
vocab. GPT-2: other targets	27.316	24.143	21.039	16.164	19.353	14.574	16.607	17.346
vocab. human: new	10.373	5.610	5.230	3.343	3.040	3.227	2.844	3.440
vocab. human: not new	15.417	19.170	16.773	18.053	14.264	14.241	16.784	14.251

Table 5: All results for HS/CN pairs.

versions	V_2	V_3	V_4	V_5	$V_{6,SBF}$	$V_{6,ARG}$	$V_{6,LAB}$	$V_{6,MIX}$
HS metrics								
HTER (all pairs)	0.399	0.497	0.405	0.189	0.424	0.301	0.385	0.444
RR	6.508	9.496	9.101	15.576	9.062	10.479	10.700	11.361
V_i vs. $cumulative$ novelty	0.757	0.697	0.624	0.463	0.618	0.436	0.582	0.606
V_i vs. V_1 novelty	0.757	0.761	0.731	0.678	0.760	0.689	0.743	0.758
V_i vs. V_{i-1} novelty	0.757	0.713	0.662	0.522	0.673	0.652	0.645	0.666
CN metrics								
HTER (all pairs)	0.527	0.383	0.351	0.312	0.329	0.342	0.382	0.341
RR	2.443	3.692	4.236	5.690	4.428	4.125	4.521	4.428
V_i vs. $cumulative$ novelty	0.814	0.788	0.760	0.737	0.743	0.743	0.741	0.745
V_i vs. V_1 novelty	0.814	0.806	0.800	0.795	0.805	0.801	0.801	0.802
V_i vs. V_{i-1} novelty	0.814	0.798	0.779	0.758	0.771	0.771	0.768	0.774

Table 6: All results for HS and CN sets alone.

HS: *Every Muslim is a potential terrorist, even if they are not born here!* **CN:** *A person's place of birth is important. Some people prefer to identify as non-Muslim, some as Muslim, and still others as Christian. I prefer to call them what they are: people.* **HS**$_{pe}$**:** Every Muslim is a potential terrorist, even if they are born here! **CN**$_{pe}$**:** A person's place of birth is not relevant, but neither is a person's religion. Some people are Muslims. I prefer to call them what they are: people, not potential terrorists. **TARGET:** `MUSLIMS`
HS: *men are more smart than women* **CN:** *It is about time women are given a chance to prove that they are intelligent as well.* **HS**$_{pe}$**:** Men are smarter than women, who should be excluded from education. **CN**$_{pe}$**:** This is not true: it is about time women are given a chance to prove that they are intelligent as well... and it is sad that they still need to prove it. **TARGET:** `WOMEN`
HS: *Gays are just a bunch of disgusting animals trying to ruin other animals' lives.* **CN:** *Being gay is not a problem, it is about orientation.*
HS$_{pe}$**:** Gays are just a bunch of disgusting animals trying to ruin other people's lives. **CN**$_{pe}$**:** Being gay is not a problem, being homophobic is. **TARGET:** `LGBT+`
HS: *Jews are responsible for the 9/11 attacks.* **CN:** *I don't think this idea is true: since the 9/11 attacks were committed by Muslims, it doesn't seem fair to blame Jews.* **HS**$_{pe}$**:** Jews are responsible for the 9/11 attacks. **CN**$_{pe}$**:** I don't think this idea is true: since the 9/11 attacks were committed by al Qaeda extremists it doesn't seem fair to blame Jews. **TARGET:** `JEWS`

Table 7: Examples of HS/CN pairs before and after post-editing with assigned target labels.