On the Impact of Hate Speech Synthetic Data on Model Fairness

Camilla Casula[1], Sara Tonelli[1]

[1]Fondazione Bruno Kessler, Trento, Italy

Abstract

Although attention has been devoted to the issue of online hate speech, some phenomena, such as ableism or ageism, are scarcely represented by existing datasets and case studies. This can lead to hate speech detection systems that do not perform well on underrepresented identity groups. Given the unprecedented capabilities of LLMs in producing high-quality data, we investigate the possibility of augmenting existing data with generative language models, reducing target imbalance. We experiment with augmenting 1,000 posts from the Measuring Hate Speech corpus, an English dataset annotated with target identity information, adding around 30,000 synthetic examples using both simple data augmentation methods and different types of generative models, comparing autoregressive and sequence-to-sequence approaches. We focus our evaluation on the performance of models on different identity groups, finding that performance can differ greatly for different targets and "simpler" data augmentation approaches can improve classification better than state-of-the-art language models.

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

Keywords

hate speech detection, synthetic data, model fairness, hate speech target

1. Introduction

Generic hate speech detection models can nowadays achieve high performance on benchmark datasets, especially for high-resource languages [1]. However, these models can still present a number of issues and weaknesses. In particular, the creation and maintenance of corpora for this task can be problematic due to the relative scarcity of hateful data online [2], the negative psychological impact on annotators [3], dataset decay and therefore reproducibility of results [4], and more.

Hate speech detection models have also been found to often have a tendency to over-rely on specific identity terms, in particular minority group mentions and other identity-related terms [5, 6, 7]. Another issue with existing datasets and systems for this task is related to the representation of identity groups that are targets of hate, which is rather unbalanced. For example, misogyny has been covered in several datasets [8, 9], while other phenomena have received much less attention, such as religious hate [10] or hate against LGBTQIA+ people [11, 12, 13]. Furthermore, phenomena such as ageism and ableism have only been marginally addressed, as shown in the survey by Yu et al. [14]. This disparity affects in turn system fairness, because offenses against less-represented targets will be classified with a lower accuracy, further impacting communities that are already marginalized [15]. By *fairness*, in this work we mean group fairness, which implies independence between model classification outputs and sensitive attributes [16].

A potential solution that has been proposed for many of the issues with hate speech detection data is the creation of synthetic data [17]. Indeed, recent research has shown it to be a promising solution [18, 19, 20, 21], albeit with mixed results [22, 23]. However, no in-depth analysis of the effects of data augmentation (DA) for less represented hate speech targets has been carried out, while it could be beneficial not only to make systems more accurate and robust, but also *fairer*, with comparable performance on hate speech targeting different demographic groups [16]. Another aspect we investigate in this work is a comparison between recent generative language models and more traditional approaches to data augmentation with regards to hate speech detection, since increasing the amount of training data with synthetic examples has been successfully exploited well before the advent of generative large language models, and can lead to improvements although these methods have a much lower computational cost [24].

In this work, we therefore address the following research questions:

(Q1) What is the impact of data augmentation on model performance for specific target identities?

(Q2) Can information about identity groups in the generation process help the creation of better and more representative synthetic examples?

(Q3) Can certain data augmentation setups enhance the performance of models on underrepresented targets, therefore improving their fairness by reducing differences in performance across different identity groups?

We aim at answering these questions through a set of experiments in which we focus on the performance of

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

✉ ccasula@fbk.eu (C. Casula); satonelli@fbk.eu (S. Tonelli)

models by target identity. In addition, we introduce two novel elements compared to previous work on generative DA: *(i)* we experiment with setups in which we exploit target identity information during generation, attempting to increase the relative representation of scarcely represented targets, with the aim of positively impacting model fairness, and *(ii)* we experiment with instruction-finetuned large language models (LLMs), which have recently been shown to be able to improve downstream task performance [25]. We also further investigate potential fairness-related weaknesses of models using the HateCheck test suite [7] combined with a manual analysis of generated examples.

2. Background

The field of hateful content detection has gained a large amount of traction in recent years, with increased effort from the research community in establishing common guidelines and benchmarks (e.g. Basile et al. [26], Zampieri et al. [27]) across different languages and targets of hate [28, 29, 11, 30].

A potential way that has been proposed to mitigate some of the issues with hate speech datasets, such as data scarcity [2] and negative psychological impact on annotators [3], is data augmentation, which could also benefit the performance of hate speech detection systems. Data augmentation refers to a family of approaches aimed at increasing the diversity of training data without collecting new samples [31]. While DA is widely used to make models more robust across many machine learning applications, it has not been as frequently adopted or researched in NLP [32, 33] until recently, with LLMs that are capable of generating realistic text [34, 35].

DA for the detection of hate speech has recently been explored using generative LLMs: Juuti et al. [36] use GPT-2 [37] to augment toxic language data in extremely low-resource scenarios. Similarly, Wullach et al. [18] and D'Sa et al. [19] successfully augment toxic language datasets using GPT-2. Fanton et al. [38] combine GPT-2 and human validation to create counter-narratives that cover multiple hate targets. More recently, Ocampo et al. [39] have applied data augmentation to increase the number of instances for the minority class in implicit and subtle examples of hate speech. Casula and Tonelli [22] show that generative data augmentation for hate speech detection using GPT-2 is in some cases challenged by a simple oversampling baseline, while Casula et al. [23] analyse the qualitative differences between original and paraphrased hate speech data. Finally, Hartvigsen et al. [20] use manually curated (through a human-in-the-loop process) prompts to generate implicitly hateful sequences with GPT-3 [40].

To our knowledge, no dedicated analyses have been carried out on the impact data augmentation can have on the performance of models for specific targets of hate, or into the exploitation of target identity information to potentially improve fully automated data augmentation processes.

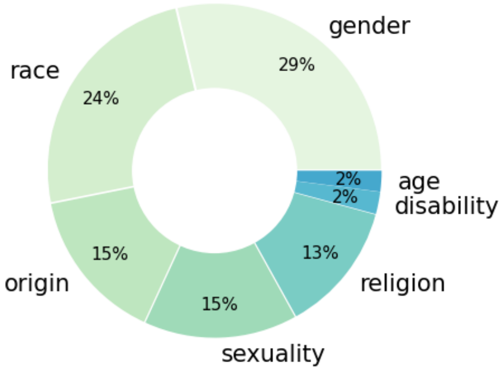

Figure 1: Identity group distribution in the MHS corpus.

3. Data

For our experiments, we use the Measuring Hate Speech (MHS) Corpus [41, 42], a dataset consisting of social media posts in English from three social media platforms (Reddit, Twitter, and YouTube). While the corpus is meant to capture different levels of hatefulness on a scale, it also includes binary hate speech labels for benchmarking purposes, which we use in our experiments.

The MHS corpus features labels regarding the binary identification of pre-specified identity groups and subgroups in texts. Importantly, this annotation is present regardless of hatefulness, resulting in target annotations even for posts containing supportive or counter-speech. In the MHS dataset [1] we find annotations for seven target identity groups: *race*, *religion*, *origin*, *gender*, *sexuality*, *age*, and *disability*. Their distribution in the data can be seen in Figure 1, which shows how the most widely studied targets of hate speech, *race* and *gender*, are also the most widely represented in the MHS corpus.

Given that the MHS corpus uses disaggregated annotations, we aggregate them so that each example has a unique label and set of targets. First, we consider each example to be about or targeting all the identity groups identified by at least half of the annotators who annotated it. Since the `hatespeech` label in the dataset can assume three values (0: *non hateful*, 1: *unclear*, 2: *hateful*), we binarize these by averaging all the annotations for

[1] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

a given post, mapping it to *hateful* if the average score is higher than 1 and to *non hateful* if it is lower.[2] After this process, we are left with 35,243 annotated posts, of which 9,046 are annotated as containing hate speech.

4. Methodology

For our experiments, we compare different generation strategies to train hate speech detection models of different sizes, aiming at assessing the impact of data augmentation based on language models on specific target identities. In order to do this, we evaluate both decoder-only and encoder-decoder models, experimenting also with their instruction-tuned counterparts. Additionally, we experiment with the inclusion of target identity information in the prompts, with the assumption that this information might lead to more varied and representative generated texts. We then use two different methods of exploiting existing information and data to generate new sequences: finetuning and few-shot prompting.

4.1. Generative Models

While most of the work on generation-based data augmentation for this task focuses on decoder-only Transformer models [22], other works have shown encoder-decoder Transformers to be potentially effective as well [43]. Since no work has been carried out on comparing decoder-only with encoder-decoder models for this type of data augmentation, we experiment with both. Then, based on work showing how instruction-tuning can improve generalization to unseen tasks [25, 44], we aim at experimenting also with instruction-finetuned models.

To favor reproducibility, we choose to only use openly available models for our experiments. We employ Llama 3.1 8B in its base and *Instruct* versions [45], OPT in its base and IML (instruction-tuned) versions [46] and T5 in its base and FLAN (instruction-tuned) versions [47, 44]. We use the 1.3B parameter version of OPT and OPT-IML and the Large version of T5 and Flan-T5 (770M), aiming at capturing in our analyses the effects of this kind of methodology with different model sizes.

4.2. Target Identity Information

In addition to performing DA with different types of models and techniques, we investigate for the first time the possibility of including target identity information both when finetuning models and when prompting them, with the hypothesis that the inclusion of this kind of information might help in generating more varied data with regards to identity group mentions for both hateful and non-hateful messages. By generating target-specific examples also for the non-hateful class, we ideally aim at implicitly contrasting identity term bias. In order to do this, we encode target identity information into the prompts given to the models in various ways.

4.3. Finetuning vs Few-Shot Prompting

A large number of works on data augmentation based on generative models rely on finetuning a model on a small set of gold data, and then generating new data with the finetuned model, encoding the label information within the text sequences in some form (e.g. Anaby-Tavor et al. [34], Kumar et al. [35]). Other works use few-shot demonstration-based prompting, in which the pretrained model is prompted with one or more sequences similar to what the model is expected to generate, with no finetuning (e.g. Hartvigsen et al. [20], Azam et al. [43], Ashida and Komachi [48]). We experiment with both strategies.

Finetuning (FT) For finetuning, we follow an approach similar to that of Anaby-Tavor et al. [34], in which a generative LLM is finetuned on annotated sequences that are concatenated with labels. At generation time, the desired label information is fed into the model, and the model is expected to generate a sequence belonging to the specified class. We discuss the details of the formatting of the label information in Section 4.4.

This method has the upside of theoretically being more likely to generate examples that are closer to the original distribution of the data to be augmented. However, this can also be a downside, if the desired effect is increasing the variety of the data. In addition, finetuning is more computationally expensive than few-shot prompting.

For models finetuned with target identity information, given that each sequence can be associated with more than one target (in cases of intersectional hate speech for instance), a different label-encoding sequence will be used to include all target identities represented in that post. An example of prompt to produce a a post about *gender* that is hateful is *Write a hateful social media post about gender.*

Few-shot prompting (FS) Following the large amount of works focusing on few-shot demonstration-based instructions, especially with instruction-finetuned models [49, 44], we also experiment with demonstration-based prompting, in which the models are shown 3 examples belonging to the desired label (and target identity, if available), and then asked to produce a new one.

With models exploiting target identity information for few-shot prompting, we associate the desired label and

[2] While we are aware this does not exploit the most novel and interesting features of the MHS dataset, the exploration of annotator (dis)agreement with regards to data augmentation is beyond the scope of this work, and is left for future research.

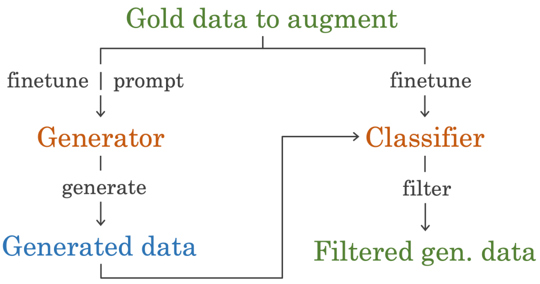

Figure 2: Generative DA pipeline.

target with 3 sequences. For instance, if the model is expected to generate a non-hateful post about gender, we select 3 sequences that are annotated in the gold data as non-hateful and about gender.

Filtering In both cases, following previous work, we perform an additional filtering step after generating, in order to try and filter out synthetic examples that are assigned an incorrect label, since label assignment is not always reliable with this type of DA [35]. The filtering step consists in feeding the generated sequences into a classifier trained on the initial non-augmented data, and only preserving sequences that are predicted by the classifier as belonging to the same label that was prompted to the generative model at generation time. An overview of the data augmentation pipeline we use is shown in Figure 2.

4.4. Prompting and Formatting

We aim at using the same type of prompting layout across experiments. We choose to use prompting sequences in natural language, given that they have been found to lead to generally more realistic generated examples for this purpose [22]. In order to find prompts in natural language that could be leveraged by our models, we consulted the FLAN corpus [25], which is part of the finetuning data of both FLAN-T5 and OPT-IML. Among the instruction templates, we find one of the CommonGen templates [50] to fit with our aims: '*Write a sentence about the following things: [concepts], [target]*'. We reformulate it to obtain a prompting sequence that reflects our application, and can be exploited by instruction-finetuned models: *Write a [∅/ hateful] social media post [∅/ about t]*, where t is a target identity category.

5. Experimental Setup

For all experiments, we simulate a setup in which we have a small amount of gold data available prior to augmentation, following previous work on data augmentation in which data scarcity is simulated to assess the effectiveness of data augmentation [34, 22]. We randomly select 1,000 examples, as we deem it a realistically small dataset size for a hate speech detection corpus on the small end of the spectrum, after looking at the hate speech dataset review by Vidgen and Derczynski [17].

Our goal is to create a larger dataset out of the starting 1,000 examples. Given that the 'natural' size of the Measuring HS dataset is 35k examples, we aim for 30k new annotated examples to use in augmentation, which will result in a 31k example dataset for each setup. We generate $\sim$ 3x the examples we need, based on the findings of Wullach et al. [18], setting the total number of generated examples for each setup to 100,000. For each setup, we prompt models to generate 50k for each label, ideally mitigating label imbalance.

Given that our focus is on different targets of hate, we aim at increasing the percentage of posts targeting or about scarcely represented minorities, ideally in order to make systems fairer, so we augment each target identity category equally. For models that rely on target identity information, out of the 50k to generate for each label, we generate 1/7 for each target (7,140 sequences).

Once we generate 100,000 examples, we feed them into a DeBERTa-v3 Large classifier [51] finetuned on the initial 1,000 gold examples, and we only preserve the examples for which the classifier label assignment matches the desired label that was in the model input at generation time. If less than 15k generated sequences pass filtering, we preserve the examples that did pass filtering, and proceed with the rest of the pipeline.

We mainly test the quality of the synthetic data extrinsically, i.e., we test its usefulness when it is used for training models, using the performance of models trained on the synthetically augmented data as a proxy for the quality of the synthetic data itself. To do this, we use synthetic data in addition to the initial available gold data for training classifiers aimed at detecting the presence of hate speech. The reasoning behind this choice is that better-quality synthetic data should lead to better performance of models trained on data augmented with it, as that is ultimately the purpose of the data in our case. Further details about our implementations, including hyperparameters and random seeds for reproducibility, are reported in Appendix B.

Baselines We implement three baselines using DeBERTa: *i)* the classifier finetuned on the starting 1k gold examples; *ii)* the same classifier finetuned on an oversampled version of the training data (repeating the initial 1k sequences until we get to 31k, the size of the augmented setups), which has been found effective even in cross-dataset scenarios [22]; and *(iii)* as a stronger baseline, we also compare all of our models with models trained on

Table 1
DeBERTa results (macro-F1 and hate-class F1) with generative DA, averaged over 5 runs $^{\pm stdev}$, overall and by target (*Ge*nder, *Ra*ce, *Or*igin, *Se*xuality, *Re*ligion, *Di*sability, and *Ag*e). n(h) = number of hateful synthetic examples preserved after filtering. *FT* stands for *fine-tuning*, while *FS* stands for *few-shot*. The *Tar* columns refers to the presence of target information in the fine-tuning or the prompting of the model.

			M-F1	h-F1	Ge h-F1	Ra h-F1	Or h-F1	Se h-F1	Re h-F1	Di h-F1	Ag h-F1	n(h)
No augmentation			$.773^{.02}$	$.652^{.03}$	$.635^{.02}$	$.696^{.04}$	$.497^{.05}$	$.756^{.03}$	$.485^{.12}$	$.698^{.03}$	$.545^{.04}$	
Oversampling			$.773^{.02}$	$.653^{.04}$	$.652^{.05}$	$.740^{.02}$	$.568^{.05}$	$.787^{.02}$	$.571^{.03}$	$.732^{.04}$	$.555^{.06}$	
EDA			$.799^{.01}$	$\mathbf{.714}^{.01}$	$\mathbf{.687}^{.01}$	$\mathbf{.771}^{.02}$	$\mathbf{.582}^{.03}$	$\mathbf{.806}^{.01}$	$\mathbf{.601}^{.02}$	$\mathbf{.799}^{.02}$	$.589^{.06}$	15k
Model		Tar										
Llama 3.1 8B	FT	Y	$.778^{.02}$	$.668^{.05}$	$.649^{.02}$	$.710^{.06}$	$.526^{.08}$	$.773^{.03}$	$.505^{.05}$	$.731^{.04}$	$.613^{.06}$	7k
		N	$.779^{.03}$	$.663^{.06}$	$.630^{.09}$	$.716^{.05}$	$.523^{.07}$	$.763^{.07}$	$.515^{.09}$	$.732^{.06}$	$.577^{.13}$	15k
	FS	Y	$\mathbf{.801}^{.01}$	$.701^{.02}$	$.678^{.03}$	$.767^{.01}$	$.578^{.05}$	$.801^{.02}$	$.589^{.04}$	$.774^{.03}$	$\mathbf{.620}^{.08}$	11k
		N	$.791^{.01}$	$.690^{.02}$	$.666^{.03}$	$.744^{.03}$	$.548^{.03}$	$.786^{.03}$	$.533^{.06}$	$.765^{.04}$	$.566^{.08}$	12k
Llama 3.1 8B Instruct	FT	Y	$.786^{.01}$	$.682^{.02}$	$.661^{.02}$	$.735^{.02}$	$.525^{.05}$	$.784^{.03}$	$.534^{.04}$	$.724^{.03}$	$.571^{.09}$	6k
		N	$.796^{.01}$	$.690^{.02}$	$.682^{.03}$	$.765^{.02}$	$.575^{.06}$	$.799^{.02}$	$.572^{.07}$	$.759^{.14}$	$.590^{.05}$	15k
	FS	Y	$.791^{.02}$	$.695^{.03}$	$.669^{.03}$	$.762^{.03}$	$.565^{.02}$	$.796^{.04}$	$.552^{.05}$	$.761^{.03}$	$.572^{.06}$	13k
		N	$.796^{.01}$	$.709^{.02}$	$.679^{.03}$	$.768^{.02}$	$.564^{.03}$	$.797^{.02}$	$.575^{.04}$	$.753^{.04}$	$.591^{.06}$	15k
OPT	FT	Y	$.783^{.00}$	$.683^{.01}$	$.653^{.03}$	$.740^{.02}$	$.556^{.05}$	$.779^{.02}$	$.535^{.07}$	$.777^{.02}$	$.587^{.06}$	0.5k
		N	$.774^{.04}$	$.652^{.07}$	$.634^{.05}$	$.707^{.06}$	$.505^{.06}$	$.738^{.10}$	$.461^{.07}$	$.690^{.11}$	$.590^{.10}$	15k
	FS	Y	$.782^{.01}$	$.691^{.02}$	$.667^{.02}$	$.750^{.02}$	$.553^{.04}$	$.790^{.01}$	$.546^{.05}$	$.791^{.02}$	$.582^{.07}$	11k
		N	$.791^{.01}$	$.700^{.01}$	$.675^{.02}$	$.758^{.02}$	$.561^{.02}$	$.791^{.02}$	$.555^{.07}$	$.776^{.03}$	$.597^{.05}$	15k
OPT IML	FT	Y	$.789^{.01}$	$.681^{.02}$	$.661^{.02}$	$.720^{.05}$	$.516^{.09}$	$.789^{.01}$	$.493^{.05}$	$.735^{.04}$	$.579^{.06}$	15k
		N	$.796^{.01}$	$.690^{.02}$	$.674^{.03}$	$.738^{.02}$	$.500^{.07}$	$.791^{.02}$	$.488^{.10}$	$.723^{.09}$	$.593^{.10}$	15k
	FS	Y	$.789^{.01}$	$.698^{.01}$	$.672^{.02}$	$.757^{.02}$	$.563^{.03}$	$.798^{.02}$	$.552^{.07}$	$.780^{.03}$	$.577^{.07}$	11k
		N	$.792^{.01}$	$.699^{.01}$	$.673^{.02}$	$.755^{.02}$	$.564^{.03}$	$.795^{.01}$	$.558^{.06}$	$.772^{.04}$	$.604^{.05}$	15k
T5	FT	Y	$.792^{.01}$	$.696^{.02}$	$.667^{.02}$	$.753^{.02}$	$.567^{.04}$	$.795^{.02}$	$.566^{.05}$	$.771^{.03}$	$.584^{.09}$	12k
		N	$.789^{.01}$	$.684^{.01}$	$.660^{.02}$	$.731^{.03}$	$.536^{.02}$	$.784^{.01}$	$.523^{.08}$	$.748^{.04}$	$.592^{.07}$	10k
	FS	Y	$.786^{.01}$	$.682^{.02}$	$.674^{.03}$	$.738^{.02}$	$.500^{.07}$	$.791^{.02}$	$.488^{.10}$	$.723^{.09}$	$.593^{.10}$	11k
		N	$.798^{.01}$	$.700^{.02}$	$.666^{.02}$	$.756^{.02}$	$.559^{.07}$	$.793^{.01}$	$.573^{.05}$	$.774^{.03}$	$.596^{.04}$	15k
FLAN T5	FT	Y	$.792^{.01}$	$.696^{.01}$	$.669^{.01}$	$.752^{.01}$	$.559^{.03}$	$.792^{.02}$	$.574^{.05}$	$.767^{.03}$	$.600^{.07}$	14k
		N	$.793^{.01}$	$.691^{.01}$	$.672^{.02}$	$.737^{.03}$	$.544^{.05}$	$.790^{.01}$	$.520^{.08}$	$.750^{.04}$	$.597^{.08}$	10k
	FS	Y	$.786^{.00}$	$.684^{.01}$	$.651^{.02}$	$.743^{.02}$	$.558^{.04}$	$.778^{.01}$	$.536^{.04}$	$.744^{.04}$	$.590^{.10}$	0.3k
		N	$.774^{.02}$	$.662^{.04}$	$.637^{.04}$	$.709^{.06}$	$.509^{.09}$	$.765^{.03}$	$.490^{.09}$	$.724^{.06}$	$.583^{.09}$	0.3k

data augmented using Easy Data Augmentation (EDA) [52]. EDA consists of four operations: synonym replacement, random insertion, random swap, and random deletion of tokens. Similarly to our other setups, we produce 30k new sequences with EDA, of which 7,500 with each operation, on the initial 1,000 examples in each fold. We then also experiment with the mixture of EDA and generative DA, in which instead of augmenting the initial gold data with 30k synthetic sequences obtained with EDA or generative DA, we randomly select 15k examples of each and concatenate them.

6. Results and Discussion

In this section we report the results of our experiments, averaged across 5 data folds using different random seeds. The performance of our baselines and models trained on generation-augmented data in terms of macro-averaged F1 score and hateful class F1 (*h-F1*)both globally and by target identity group is reported in Table 1. All models are tested on a held-out portion of the gold data from the MHS corpus.

Considering simply the *no augmentation* baseline, it is clear that performance can vary greatly across target groups, with up to 27% hate-F1 differences between them. In particular, the model appears to struggle with posts about *origin* (Or), *religion* (Re), and *age* (Ag), while, although underrepresented compared to other target groups, posts about *disability* (Di) tend to be classified more accurately on average. This suggests that performance might also be influenced by factors other than the representation of targets in the dataset, such as how broad a target category is or how much variation there is within it. For instance, *origin* can include any type of discrimination based on geographical origin, potentially making it harder to generalize for, and *religion* as a category encompasses any type of religious discourse, in spite of each religion being targeted through specific offense types [10]. This makes classification challenging, especially for systems that rely primarily on lexical features.

Most of the models trained on generation-augmented data outperform the *no augmentation* baseline across targets, with different improvements based on target identity group (*origin*, *religion*, and *age* in particular). Strikingly, however, EDA performs better than all generation-based DA configurations, regardless of prompting type or access to target information, for all targets but *age*.

We hypothesize EDA is effective because small perturbations can make models more robust, especially with regard to the *hateful* class, while generative models do increase performance, but they are also more likely to inject noise.

The impact of finetuning vs. few-shot prompting seems model-dependent, with differences across models also regarding the impact of target information. Interestingly, the amount of synthetic examples labeled as *hateful* that pass filtering does not appear to be linked with better performances of models trained on synthetic data.

7. Qualitative Analysis

In this section, we look into the synthetically generated texts and the models trained on them from a qualitative point of view. First we carry out a manual annotation on the generated texts. Then, we turn to the HateCheck test suite [7], which includes examples aimed at exploring the weaknesses of hate speech models, especially their out-of-distribution generalization, again focusing on performance by target. HateCheck targets are in some cases more specific than those present in our dataset, thus providing a complementary view on our models' performance.

7.1. Manual Annotation

A total of 1,120 generated texts filtered with DeBERTa were annotated by two annotators with a background in linguistics and experience in hate speech research. For each combination of finetuning/prompting/target presence for each model, they annotated 70 examples, evenly distributed across labels and, where available, targets. The examples were annotated according to *label correctness*, *target category correctness* (where available), and *realism.*

For the examples generated *without access to target information*, the *target* dimension was not annotated. Consider for example the following sentence, generated giving 'age' as target information: *'F*ckin white men are trashy like a muthaf*cker'.* In this case, Label would be '*hateful*', Realism would be '*Yes*' but Target would be '*No*', because the target identity category of the generated example is 'race' and not 'age'.

Table 2
Generated texts labeled as correct by human annotators in terms of labels, target categories, and realism. N/A refers to cases in which all of the generated texts were nonsensical (0% realistic), with impossible assignment of labels or categories.

Model		Tar	Label	Target	Realism
Llama 3.1 8B	FT	Y	98%	72%	89%
		N	87%	/	86%
	FS	Y	93%	53%	86%
		N	90%	/	84%
Llama 3.1 8B Inst.	FT	Y	87%	66%	79%
		N	87%	/	73%
	FS	Y	89%	61%	81%
		N	83%	/	79%
OPT	FT	Y	93%	63%	66%
		N	N/A	/	0%
	FS	Y	90%	39%	83%
		N	81%	/	70%
OPT-IML	FT	Y	96%	53%	66%
		N	N/A	/	0%
	FS	Y	90%	57%	79%
		N	81%	/	73%
T5	FT	Y	83%	59%	80%
		N	74%	/	30%
	FS	Y	N/A	N/A	0%
		N	N/A	/	0%
Flan-T5	FT	Y	94%	66%	81%
		N	74%	/	41%
	FS	Y	89%	36%	84%
		N	87%	/	86%

Inter-annotator agreement was calculated using Krippendorff's alpha on 10% of the manually analyzed data (112 examples). The annotators showed moderate agreement with regards to label correctness (α = 0.76), while the scores were higher for category correctness (α = 0.83) and realism (α = 0.82).

The results of the manual analysis are reported in Table 2. In most cases, the addition of target information results in more realistic texts and, in general, more accurate label assignment. However, this is not directly associated with improved model performance from augmented data. In addition, the rate of realistic texts and the accuracy of the identity categories are still somewhat low compared to the correctness of label assignment, showing that the generative models we tested might have difficulties dealing with more than one type of constraint/instruction. Indeed, while few-shot (FS) approaches sometimes lead

Table 3
DeBERTa results on HateCheck (hate-class F_1) by target identity, averaged over 5 runs $^{\pm stdev}$. *p.* is an abbreviation for *people*, while *disab* stands for *people with disabilities*. *FT* stands for *fine-tuning*, while *FS* stands for *few-shot*. The *Tar* columns refers to the presence of target information in the fine-tuning or the prompting of the model.

			Women	Trans p.	Gay p.	Black p.	Disab.	Muslims	Immigrants
No Augmentation			$.142^{.05}$	$.101^{.03}$	$.252^{.06}$	$.216^{.07}$	$.113^{.04}$	$.147^{.04}$	$.109^{.01}$
EDA			$\mathbf{.400}^{.04}$	$\mathbf{.485}^{.09}$	$\mathbf{.590}^{.06}$	$\mathbf{.643}^{.09}$	$\mathbf{.463}^{.11}$	$\mathbf{.546}^{.13}$	$\mathbf{.420}^{.06}$
Model		Target							
Llama 3.1 8B	FT	Y	$.240^{.15}$	$.166^{.10}$	$.331^{.10}$	$.300^{.16}$	$.189^{.16}$	$.212^{.15}$	$.173^{.14}$
		N	$.126^{.11}$	$.084^{.08}$	$.211^{.13}$	$.212^{.13}$	$.096^{.08}$	$.123^{.09}$	$.080^{.06}$
	FS	Y	$.286^{.08}$	$.203^{.07}$	$.371^{.14}$	$.433^{.10}$	$.232^{.05}$	$.419^{.07}$	$.287^{.05}$
		N	$.239^{.06}$	$.184^{.08}$	$.294^{.07}$	$.389^{.11}$	$.223^{.08}$	$.285^{.12}$	$.222^{.09}$
Llama 3.1 8B Instruct	FT	Y	$.206^{.10}$	$.142^{.09}$	$.329^{.15}$	$.293^{.19}$	$.161^{.09}$	$.223^{.17}$	$.166^{.14}$
		N	$.178^{.08}$	$.137^{.07}$	$.270^{.07}$	$.260^{.09}$	$.142^{.06}$	$.200^{.12}$	$.134^{.09}$
	FS	Y	$.224^{.11}$	$.205^{.09}$	$.315^{.08}$	$.332^{.06}$	$.203^{.12}$	$.245^{.07}$	$.152^{.10}$
		N	$.196^{.06}$	$.195^{.07}$	$.336^{.12}$	$.322^{.09}$	$.212^{.08}$	$.215^{.11}$	$.148^{.09}$
OPT	FT	Y	$.233^{.08}$	$.197^{.09}$	$.340^{.13}$	$.327^{.11}$	$.253^{.08}$	$.253^{.09}$	$.212^{.09}$
		N	$.109^{.05}$	$.057^{.03}$	$.167^{.06}$	$.162^{.06}$	$.086^{.03}$	$.092^{.04}$	$.067^{.03}$
	FS	Y	$.283^{.10}$	$.237^{.12}$	$.424^{.13}$	$.457^{.13}$	$.254^{.09}$	$.352^{.10}$	$.261^{.08}$
		N	$.249^{.02}$	$.218^{.07}$	$.383^{.10}$	$.423^{.10}$	$.235^{.04}$	$.278^{.08}$	$.234^{.10}$
OPT- IML	FT	Y	$.189^{.07}$	$.127^{.05}$	$.239^{.06}$	$.201^{.07}$	$.151^{.08}$	$.162^{.07}$	$.126^{.07}$
		N	$.124^{.06}$	$.057^{.03}$	$.155^{.04}$	$.137^{.04}$	$.082^{.04}$	$.086^{.05}$	$.058^{.04}$
	FS	Y	$.297^{.07}$	$.234^{.07}$	$.378^{.07}$	$.406^{.13}$	$.232^{.08}$	$.366^{.05}$	$.244^{.06}$
		N	$.238^{.12}$	$.209^{.13}$	$.366^{.17}$	$.403^{.18}$	$.232^{.11}$	$.273^{.14}$	$.194^{.10}$
T5	FT	Y	$.259^{.10}$	$.240^{.11}$	$.409^{.10}$	$.428^{.17}$	$.276^{.12}$	$.385^{.12}$	$.273^{.10}$
		N	$.148^{.08}$	$.106^{.06}$	$.275^{.13}$	$.260^{.12}$	$.125^{.06}$	$.147^{.05}$	$.111^{.04}$
	FS	Y	$.150^{.08}$	$.093^{.05}$	$.220^{.07}$	$.231^{.15}$	$.111^{.06}$	$.200^{.10}$	$.128^{.07}$
		N	$.219^{.05}$	$.137^{.16}$	$.289^{.08}$	$.282^{.09}$	$.157^{.03}$	$.229^{.06}$	$.177^{.05}$
Flan-T5	FT	Y	$.185^{.08}$	$.120^{.07}$	$.250^{.10}$	$.268^{.15}$	$.154^{.09}$	$.254^{.14}$	$.178^{.09}$
		N	$.143^{.04}$	$.076^{.04}$	$.218^{.03}$	$.202^{.06}$	$.114^{.02}$	$.146^{.05}$	$.098^{.03}$
	FS	Y	$.252^{.08}$	$.188^{.08}$	$.313^{.09}$	$.346^{.14}$	$.210^{.07}$	$.284^{.08}$	$.206^{.07}$
		N	$.248^{.11}$	$.198^{.10}$	$.319^{.11}$	$.326^{.17}$	$.187^{.09}$	$.252^{.11}$	$.196^{.10}$

to more realistic generated sequences, this often entails lower label or category correctness and vice versa.

7.2. HateCheck

We perform a second qualitative analysis using the HateCheck test suite [7], a collection of functional testing examples that enable targeted diagnostic insights of hate speech detection models. All HateCheck test cases mention a specific target identity, to allow the exploration of unintended biases against different target groups. However, the target groups used in HateCheck do not fully overlap with the target identity groups in the MHS corpus (Figure 1). The target identities in HateCheck are: women (which would fall under *gender* in MHS), trans people (*gender* in MHS), gay people (*sexuality* in MHS), black people (*race* in MHS), disabled people (*disability* in MHS), Muslims (*religion* in MHS) and immigrants (*origin* in MHS). The *age* category is present in MHS corpus and entirely missing in HateCheck.

We report in Table 3 the results obtained by the models trained on augmented data on HateCheck in terms of hate-class F_1 score, divided by target identity group according to the identity categories used in HateCheck. Overall, we can notice significant improvements over the *no augmentation* baseline across different setups and target identities, although these improvements are variable, and again overshadowed by Easy Data Augmentation, resulting in more evenly distributed scores across all targets.

Interestingly, however, even overlapping targets between the MHS dataset and HateCheck show different improvements across generative DA setups. For instance, the HateCheck target *people with disabilities* apparently benefits from the inclusion of target information more than the *Disability* category in the MHS corpus, suggesting that the inclusion of identity group information could affect performances differently in and out of domain.

Furthermore, HateCheck test instances are manually created starting from templates that include specific mentions of identity-group terms, which could lead to sentences that are very different from 'real' examples that could be found on social media. For instance, one of the hateful templates used in HateCheck is '*I hate [IDENTITY]*', in which *[IDENTITY]* is substituted with *women*, *disabled people*, and so on for all of the targets reported above. This can result in somewhat stereotypical examples, while hateful posts in the MHS corpus often contain slurs and offensive terms instead of more neutral names to refer to people belonging to a certain identity group.

8. Conclusions

We have investigated the impact of data augmentation with generative models on specific targets of hate, experimenting with instruction-finetuned models and the addition of target information when generating new sequences. Overall, it appears that DA methods have different types of impact on different targets, but they can improve performance even for scarcely represented identity categories (Q1). However, we observed that generative data augmentation alone is not as strong as simpler methods such as EDA.

Through a qualitative analysis, we also emphasized the fact that including target information when generating synthetic examples can facilitate the creation of examples that are more realistic and exhibit more correct label assignments (Q2), although further work could investigate why these characteristics do not directly correlate with downstream task performance.

Overall, our analysis shows that there is potential in data augmentation with regards to model group fairness (Q3), implying independence between model classification output and sensitive attributes [16]. However, although potentially useful, this type of DA can still lead to unpredictable results, and it is not guaranteed to always improve the performance of models across all identity groups with regards to hate speech. We plan to further explore this research direction in the future, considering also intersectionality and more specific targets (e.g. groups such as *trans women* rather than the *gender* category). In addition, we worked on English data because of the availability of the Measuring Hate Speech corpus, which was large enough to perform our DA experiments and presented the kind of fine-grained target annotation required in our study. However, we are aware that DA would benefit more classification with lower-resourced languages, so we plan to work on different languages in the future.

In summary, we show that data augmentation with generative language models can be beneficial, even when using only openly available models. However, given their high computational costs, alternatives like EDA could be considered if limited resources are available, because they can still yield performance improvements compared to a low-resource setting. Again, there seems to be no one-fits-all solution or approach to generation or data augmentation in this kind of scenario.

We acknowledge that data augmentation techniques may be used also for malicious purposes, for example to create thousands of hateful examples with the goal of hurting the same groups that we want to support. Because of this, we provide all the necessary details for the reproduction of our results, but we do not plan to openly release the code or to upload the generated data produced by our experiments, especially in order to avoid it being crawled and ending up in the training data of LLMs in the future. We are, however, open to sharing the data with other researchers who might be interested.

Acknowledgments

This work was funded by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS).

References

[1] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447.

[2] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.

[3] M. J. Riedl, G. M. Masullo, K. N. Whipple, The downsides of digital labor: Exploring the toll incivility takes on online comment moderators, Computers in Human Behavior 107 (2020) 106262.

[4] F. Klubicka, R. Fernández, Examining a hate speech corpus for hate speech detection and popularity prediction, in: Proceedings of 4REAL Workshop - Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language, 2018.

[5] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and Mitigating Unintended Bias

in Text Classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, New Orleans LA USA, 2018, pp. 67–73. URL: https://dl.acm.org/doi/10.1145/3278721.3278729. doi:10.1145/3278721.3278729.

[6] B. Kennedy, X. Jin, A. Mostafazadeh Davani, M. Dehghani, X. Ren, Contextualizing hate speech classifiers with post-hoc explanation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5435–5442. URL: https://aclanthology.org/2020.acl-main.483. doi:10.18653/v1/2020.acl-main.483.

[7] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, HateCheck: Functional tests for hate speech detection models, 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), Association for Computational Linguistics, Online, 2021, pp. 41–58. URL: https://aclanthology.org/2021.acl-long.4. doi:10.18653/v1/2021.acl-long.4.

[8] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 158–168. URL: https://aclanthology.org/2020.trac-1.25.

[9] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An Expert Annotated Dataset for the Detection of Online Misogyny, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1336–1350.

[10] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128.

[11] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, 2021. arXiv:2109.00227.

[12] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 16–24.

[13] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[14] Z. Yu, I. Sen, D. Assenmacher, M. Samory, L. Fröhling, C. Dahn, D. Nozza, C. Wagner, The unseen targets of hate: A systematic review of hateful communication datasets, Social Science Computer Review (2024) 08944393241258771. doi:10.1177/08944393241258771.

[15] Z. Talat, J. Bingel, I. Augenstein, Disembodied machine learning: On the illusion of objectivity in nlp, ArXiv abs/2101.11974 (2021).

[16] J. Anthis, K. Lum, M. Ekstrand, A. Feller, A. D'Amour, C. Tan, The Impossibility of Fair LLMs, 2024. URL: http://arxiv.org/abs/2406.03198. doi:10.48550/arXiv.2406.03198, arXiv:2406.03198 [cs, stat].

[17] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (2020) e0243300. doi:10.1371/journal.pone.0243300.

[18] T. Wullach, A. Adler, E. Minkov, Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4699–4705. URL: https://aclanthology.org/2021.findings-emnlp.402. doi:10.18653/v1/2021.findings-emnlp.402.

[19] A. G. D'Sa, I. Illina, D. Fohr, D. Klakow, D. Ruiter, Exploring Conditional Language Model Based Data Augmentation Approaches for Hate Speech&#xa0;Classification, in: Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, pp. 135–146. doi:10.1007/978-3-030-83527-9_12.

[20] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326. URL: https://aclanthology.org/2022.acl-long.234. doi:10.18653/v1/2022.acl-long.234.

[21] C. Casula, E. Leonardelli, S. Tonelli, Don't augment, rewrite? assessing abusive language detection with synthetic data, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association

for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11240–11247. URL: https://aclanthology.org/2024.findings-acl.669/. doi:10.18653/v1/2024.findings-acl.669.

[22] C. Casula, S. Tonelli, Generation-based data augmentation for offensive language detection: Is it worth it?, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3359–3377. URL: https://aclanthology.org/2023.eacl-main.244.

[23] C. Casula, S. Vecellio Salto, A. Ramponi, S. Tonelli, Delving into qualitative implications of synthetic data for hate speech detection, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 19709–19726. URL: https://aclanthology.org/2024.emnlp-main.1099/. doi:10.18653/v1/2024.emnlp-main.1099.

[24] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An Empirical Survey of Data Augmentation for Limited Data Learning in NLP, Transactions of the Association for Computational Linguistics 11 (2023) 191–211. URL: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00542/2074871/tacl_a_00542.pdf.

[25] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[26] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. doi:10.18653/v1/S19-2007.

[27] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. doi:10.18653/v1/S19-2010.

[28] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D'Errico, Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP, Information Processing and Management 60 (2023) 103118. URL: https://www.sciencedirect.com/science/article/pii/S0306457322002199. doi:https://doi.org/10.1016/j.ipm.2022.103118.

[29] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188. doi:10.18653/v1/2020.semeval-1.188.

[30] E. Leonardelli, C. Casula, S. Vecellio Salto, J. E. Bak, E. Muratore, A. Kołos, T. Louf, S. Tonelli, MuLTa-Telegram: A Fine-Grained Italian and Polish Dataset for Hate Speech and Target Detection, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[31] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.

[32] L. F. A. O. Pellicer, T. M. Ferreira, A. H. R. Costa, Data augmentation techniques in natural language processing, Applied Soft Computing 132 (2023) 109803. doi:10.1016/j.asoc.2022.109803.

[33] M. Bayer, M.-A. Kaufhold, C. Reuter, A Survey on Data Augmentation for Text Classification, ACM Computing Surveys 55 (2022) 146:1–146:39. doi:10.1145/3544558.

[34] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do Not Have Enough Data? Deep Learning to the Rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390. doi:10.1609/aaai.v34i05.6233.

[35] V. Kumar, A. Choudhary, E. Cho, Data augmentation using pre-trained transformer models, in: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, Association for Computational Linguistics, Suzhou, China, 2020, pp. 18–26. URL: https://aclanthology.org/2020.lifelongnlp-1.3.

[36] M. Juuti, T. Gröndahl, A. Flanagan, N. Asokan, A little goes a long way: Improving toxic language classification despite data scarcity, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2991–3009. doi:10.18653/v1/2020.findings-emnlp.269.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei,

I. Sutskever, Language Models are Unsupervised Multitask Learners, 2019.

[38] M. Fanton, H. Bonaldi, S. S. Tekiroğlu, M. Guerini, 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), Association for Computational Linguistics, Online, 2021, pp. 3226–3240. doi:10.18653/v1/2021.acl-long.250.

[39] N. Ocampo, E. Sviridova, E. Cabrio, S. Villata, An in-depth analysis of implicit and subtle hate speech messages, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1997–2013. URL: https://aclanthology.org/2023.eacl-main.147.

[40] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, arXiv:2005.14165 [cs] (2020). arXiv:2005.14165.

[41] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application, 2020. URL: http://arxiv.org/abs/2009.10277. doi:10.48550/arXiv.2009.10277, arXiv:2009.10277 [cs].

[42] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 83–94. URL: https://aclanthology.org/2022.nlperspectives-1.11.

[43] U. Azam, H. Rizwan, A. Karim, Exploring data augmentation strategies for hate speech detection in Roman Urdu, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4523–4531. URL: https://aclanthology.org/2022.lrec-1.481.

[44] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[45] M. A. Llama Team, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[46] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, 2022. arXiv:2205.01068.

[47] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[48] M. Ashida, M. Komachi, Towards automatic generation of messages countering online hate speech and microaggressions, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 11–23. URL: https://aclanthology.org/2022.woah-1.2. doi:10.18653/v1/2022.woah-1.2.

[49] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, et al., Opt-iml: Scaling language model instruction meta learning through the lens of generalization, 2022. arXiv:2212.12017.

[50] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, X. Ren, CommonGen: A constrained text generation challenge for generative commonsense reasoning, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1823–1840. URL: https://aclanthology.org/2020.findings-emnlp.165. doi:10.18653/v1/2020.findings-emnlp.165.

[51] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. arXiv:2111.09543.

[52] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/

v1/D19-1670.

A. Prompting Examples

Below are examples the sequences and prompts used for training and prompting our models.

FT-target *Write a (hateful) social media post about {target}: {text}*

FT-no target *Write a (hateful) social media post: {text}*

FS-target *Write a (hateful) social media post about {target}: {text}* [...]
Write a (hateful) social media post about {target}: {text}

FS-no target *Write a (hateful) social media post: {text}* [...]
Write a (hateful) social media post: {text}

The values used for 'target' are the identity group names in the MHS dataset, reported in Sec. 3.

B. Hyperparameters and Reproducibility

For all of our experiments, we employ the HuggingFace Python library. All the hyperparameters we use that are not specified in this section are the default ones from their `TrainingArguments` class. The classifiers we use as baselines and for filtering are trained on 5 epochs.

We finetune all generative models with batch 16 and $LR = 1e-3$. For generation, we set *top-p*=0.9 and min and max lengths of generated sequences to 5 and 150 tokens respectively. Finally, we avoid repeating 4-grams. All the classifiers that are trained on augmented data are trained for 3 epochs with batch size 16 and LR $5e-6$. In this case, at the end of training, we preserve the model from the epoch with the lowest evaluation cross-entropy loss.

The random seeds we used for shuffling, subsampling the gold data, and initializing both generative and classification models are 522, 97, 709, 16, and 42. These were chosen randomly. Finetuning of all classifiers and generative models, including baselines and models trained on augmented data, took 70 hours, of which 55 on a Nvidia V100 GPU and 15 on a Nvidia A40. Inference time for generating all of the sequences (a total of 8 million generated texts) took $\sim$400 hours total.