Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection

Camilla Casula [☸☾] **Sebastiano Vecellio Salto** [☸☾] **Alan Ramponi** [☸] **Sara Tonelli** [☸]
{ccasula, svecelliosalto, alramponi, satonelli}@fbk.eu
[☸] Fondazione Bruno Kessler, Italy
[☾] University of Trento, Italy

Abstract

The use of synthetic data for training models for a variety of NLP tasks is now widespread. However, previous work reports mixed results with regards to its effectiveness on highly subjective tasks such as hate speech detection. In this paper, we present an in-depth qualitative analysis of the potential and specific pitfalls of synthetic data for hate speech detection in English, with 3,500 manually annotated examples. We show that, across different models, synthetic data created through paraphrasing gold texts can improve out-of-distribution robustness from a computational standpoint. However, this comes at a cost: synthetic data fails to reliably reflect the characteristics of real-world data on a number of linguistic dimensions, it results in drastically different class distributions, and it heavily reduces the representation of both specific identity groups and intersectional hate.

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

1 Introduction

Recent advancements in generative Large Language Models (LLMs), with models having the potential to quickly produce large amounts of textual data, have resulted in a number of works on synthetic data generation in the NLP community (Feng et al., 2021; Chen et al., 2023; Li et al., 2023). Indeed, synthetic data may mitigate issues related to data scarcity, minimizing the need to collect real data and, in some cases, even to manually annotate it. Beside the advantages in terms of effort, synthetic data could comply better with privacy regulations, replacing real data with *realistic* data that can be freely shared.

In the light of this potential, recent works have tried to identify the settings and tasks in which data augmentation could be successfully employed (Chen et al., 2023). For example, Li et al. (2023) showed that classifiers trained with real data generally outperform those trained using only synthetic data, especially when the task is subjective, whereas Pendzel et al. (2023) found that synthetic data can increase cross-dataset performance. Beside extrinsic evaluations, however, little attention has been paid to the advantages and risks of employing synthetic data in sensitive tasks like hate speech detection, with the few existing efforts reporting mixed results in terms of performance (Casula and Tonelli, 2023).

In this paper, we address a scenario in which *one may need to perform hate speech detection on unseen data, and they would like to exploit the potential of both generative LLMs and existing hate speech datasets*: What advantages can synthetic data offer in this respect? What are the risks associated with using LLMs for this type of application? Could generated data amplify bias or harm? As a first exploration in this direction, we focus on hate speech detection in English, i.e., a high-resource language for which several LLMs and hate speech datasets are already available (Poletto et al., 2021).

In this work, we augment an existing English hate speech dataset, with the goal of evaluating whether, on unseen data from a different distribution, training on paraphrased data is better than using original hate speech data. We couple this evaluation with a thorough manual qualitative analysis of the generated data, assessing fluency, grammaticality and 'artificiality'. Given that biases may affect specific targets of hate differently (Sap et al., 2019, 2020), we also devote particular attention to a per-target analysis, showing the effects of the usage of LLMs to produce synthetic data on target identity distribution, and subsequently its impact on fairness.

Since generated data is increasingly being used even for sensitive applications (Ghanadian et al., 2024), it is important that also the NLP community critically addresses the impact of synthetic data including ethical risks, along the lines of similar

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19709–19726

discussions in other research communities (Whitney and Norman, 2024). Our work is an initial contribution in this direction.

2 Related Work

In the context of hate speech detection, data augmentation and synthetic data have been proposed as means to mitigate many issues in datasets (Vidgen and Derczynski, 2020; Wullach et al., 2021; Hartvigsen et al., 2022). Those include dataset decay or obsolescence and their impact on reproducibility (Klubicka and Fernández, 2018), the over-reliance of models on specific lexical items such as identity mentions (Dixon et al., 2018; Kennedy et al., 2020a; Röttger et al., 2021; De la Peña Sarracén and Rosso, 2023), and the psychological impact on annotators (Riedl et al., 2020).

The representation of minority identity groups is another issue in hate speech detection literature, with targets that have been covered extensively such as race and gender-related hate (Bhattacharya et al., 2020; Zeinert et al., 2021; Guest et al., 2021; Bosco et al., 2023), while other phenomena and targets have received less attention, such as religious hate (Ramponi et al., 2022) or hate against the LGBTQIA+ community (Chakravarthi et al., 2021; Locatelli et al., 2023).

Synthetic data-based methods have been found effective for a number of NLP tasks (Feng et al., 2021; Chen et al., 2023), with models trained on synthetic data achieving similar or, in some cases, better performance than models trained on gold data (Casula et al., 2024). Whitney and Norman (2024) categorize *synthetic data* into two separate categories, based on how derivative the data is with respect to a real-world dataset. *Generated data* refers both to an ideally 'novel' output[1] that is produced by a generative model, while *augmented data* refers to any real-world data instance that was modified in some way, e.g., via perturbations such as synonym replacement or random word deletion (Wei and Zou, 2019). Given that previous work has shown that direct generation may not work well in all scenarios for hate speech detection and in general on subjective tasks (Casula and Tonelli, 2023; Li et al., 2023), and since this is not a low-resource scenario, we frame our synthetic data creation process as a sort of middle ground between these, along the line of Casula et al. (2024): paraphrasing original real-world data rather than generating text sequences *ex-novo*. Our rationale for this choice is potentially preserving similar content to the original real-world data, while conceivably increasing the lexical variety of the data, which are typical desiderata in synthetic data approaches for this kind of task (Wullach et al., 2021).

[1]Synthetic data can hardly be entirely novel, as it is based on representations of real data (Whitney and Norman, 2024).

3 Gold Data

Most work frames synthetic data creation as a data augmentation task in a low-resource setup, i.e., by starting from a small sample of gold data. For hate speech detection, however, there exist datasets in many languages, also in the light of the shared tasks that have been organized over the years (Zampieri et al., 2019, 2020). This makes the low-resource assumption unrealistic for languages such as English. Since our study focuses on English, we devise an experimental setup that allows us to leverage existing hate speech resources by casting data augmentation as *paraphrasing* rather than as zero–shot generation. This allows us also to potentially mitigate effects related to model *alignment*, with LLMs often being programmatically blocked in generating hateful messages from scratch.

For investigating the effects that synthetic data can have on hate speech detection, we choose **Measuring Hate Speech** (MHS; Kennedy et al., 2020b; Sachdeva et al., 2022) as our corpus to paraphrase, since it covers different target identity categories. MHS includes posts from three online platforms (i.e., Twitter, Youtube, and Reddit) and contains annotations not only regarding the presence of hate speech, but also about any target identities that are mentioned in the text, regardless of whether it contains hate speech or not. Since we focus on the binary classification of hate speech, we use the *hate speech* label rather than the continuous hate speech scores. The *hate speech* label in the MHS dataset can take on three values (0: *non hateful*, 1: *unclear*, 2: *hateful*). Given that the dataset is released in disaggregated form, we manually average all the annotations for a given post, mapping the post to the *hateful* label if the average score for hate speech of that text is above 1, and to *non hateful* if it is lower. We also aggregate the information of target identities, assigning the presence of a specific target identity if at least half of the original annotators for a given post marked that identity as present. After this process, 35,242 annotated posts remained, of which 9,046 annotated as containing hate speech

(∼26%). We select 10% of the dataset as test set (3,524), 5% of the remaining examples as development data (1,586 examples) and the remaining texts as the training data (30,132 posts).

We use two more datasets for testing the out-of-distribution generalization of classifiers. First, we test our models on the **Multi-Domain Agreement** dataset (MDA; Leonardelli et al., 2021), which contains Twitter posts annotated for offensive content spanning across three main topics: the Black Lives Matter movement, Covid-19, and the 2020 US elections. For MDA, we use the default test data split (3,057 examples). Second, we test our models on the **HateCheck** dataset (Röttger et al., 2021), which contains 3,727 adversarial examples tailored at finding weaknesses of hate speech detection models.

4 Synthetic Data Generation

To be able to analyze the extrinsic impact on performance and the intrinsic characteristics of synthetic data for hate speech detection, we first artificially create training data.

Previous works focusing on synthetic data for hate speech and abusive content-related tasks have experimented with task-specific decoding (Hartvigsen et al., 2022), pipelines including humans in the loop for validating generated texts (Fanton et al., 2021; Chung et al., 2023), or fine-tuning generative large language models on real world data (Anaby-Tavor et al., 2020; Kumar et al., 2020). However, the growing performance of generative LLMs makes it possible to use them in numerous tasks without any fine-tuning (Wei et al., 2022). Because of this, we opt for a synthetic data creation setup in which we prompt LLMs to paraphrase the original texts. We expect the output text to *i)* be similar to the original social media post, *ii)* reflect the same hate speech label, and *iii)* preserve roughly the same meaning and topic. We analyze these aspects in our human evaluation in Section 6.

Our synthetic data creation pipeline consists of two steps. First, we prompt the models (Section 4.1) to obtain synthetic versions of the real data in the MHS corpus, creating one artificial counterpart for each example in the dataset. After extracting the paraphrased text from the model output, we perform two additional filtering steps on the synthetic sequences (Section 4.2).

Models We experiment with three instruction-based generative LLMs through the HuggingFace library (Wolf et al., 2020): `Llama-2 Chat 7B` (Touvron et al., 2023), `Mistral 7B Instruct v0.2` (Jiang et al., 2023), and `Mixtral 8x7B Instruct v0.1` (Jiang et al., 2024). We only use freely available and widely used models for our experiments, to favor comparability and reproducibility. We report model hyperparameters in Appendix A.

4.1 Prompting

We frame synthetic data creation as paraphrasing, as it is a common task in instruction tuning datasets that are widely used for training LLMs (Wang et al., 2022; Wei et al., 2022) and thus it does not require fine-tuning or detailed prompting. Given a text, we prompt the models with the following template:

```
Paraphrase this text: "{text}"

Paraphrased text: "
```

For `Mistral` and `Mixtral`, the template is preceded and followed by the `[INST]` and `[/INST]` tags. We then extract, using a regular expression, the first text sequence after 'Paraphrased text:' that is between inverted commas in the model output.

4.2 Filtering

We observe that, in a limited number of cases, synthetic examples are nearly identical to the original text they (should) paraphrase. We thus carry out fuzzy matching using the `thefuzz` library[2] to discard sequences that are (almost) verbatim copies of the original gold data. After some manual checks, we set the similarity threshold for discarding sequences that are too similar to 75.

In addition, a number of works on data augmentation and creation of synthetic data for this task employ a further *filtering* step (e.g., Wullach et al. (2021); Casula and Tonelli (2023)), in which the generated sequences are re-labeled using a classifier (*classifier filtering* from now on) to increase the chance that the label assignment of the synthetic texts is correct.

We aim at exploring the impact of this step, so we divide our experimental setups into:

- **No classifier filtering**, in which we preserve all synthetically created texts that passed the fuzzy matching step;
- **Classifier filtering**, in which we discard all the synthetic examples for which a classifier

[2] `pypi.org/project/thefuzz`

				Test data			
				MHS		MDA	HateCheck
		n(train)	*% hateful*	M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		30,132	26%	**.811** $\pm.004$	**.718** $\pm.008$	.507 $\pm.027$	.386 $\pm.026$
Gen. Model	Filter						
`Llama-2 Chat 7B`	No	28,289	26%	.769 $\pm.004$	.680 $\pm.003$	.675 $\pm.009$	.603 $\pm.021$
	Yes	20,187	2%	.805 $\pm.002$	.715 $\pm.002$	.539 $\pm.008$	.346 $\pm.009$
`Mistral 7B Instruct`	No	29,344	26%	.772 $\pm.004$	.686 $\pm.003$	.684 $\pm.007$	**.665** $\pm.017$
	Yes	22,483	4%	.808 $\pm.003$	.716 $\pm.004$	.526 $\pm.011$	.371 $\pm.012$
`Mixtral 8x7B Instruct`	No	29,351	26%	.754 $\pm.004$	.670 $\pm.003$	**.687** $\pm.005$	**.665** $\pm.005$
	Yes	22,370	3%	.802 $\pm.002$	.706 $\pm.003$	.525 $\pm.016$	.364 $\pm.012$

Table 1: Results of `RoBERTa Large` models trained on synthetic data only (average of 5 runs $\pm$ stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded. *Filter:Yes* means that *classifier filtering* was applied.

trained on gold data predicts a different label from the one that was assigned to the gold example the synthetic text derives from.

5 Extrinsic Evaluation

We analyze the *extrinsic* impact of synthetic data by fine-tuning classifiers on both artificial and original data. This analysis contextualizes the main contribution of this work, namely the *intrinsic* evaluation of synthetic data (Section 6), and it is aimed at addressing the following question: *What is the quantitative usefulness of synthetic data for the downstream task of hate speech detection?*

For our experiments, we use three pre-trained classifiers: `RoBERTa Large`, `RoBERTa Base` (Liu et al., 2019), and `DeBERTa v3 Base` (He et al., 2020). We compare the performance of a model trained on original gold data with the performance of the same model trained on synthetic data only and with that of a model trained on both synthetic and gold data, in order to assess how effectively the synthetic data can mimic the gold training data. For brevity, we report `Roberta Large` results in this section, since our findings are reflected across all classifiers.[3] While the classifiers are always trained on data (original or paraphrased) from MHS, they are tested on all datasets detailed in Section 3, in order to assess both their in-distribution and their out-of-distribution performance. The metrics we use for evaluating classifiers are macro-F_1 and minority class (*hate*) F_1. Details of the model implementations are reported in Appendix A.

[3]The performance of `RoBERTa Base` and `DeBERTa Base` are reported in Appendix B.

Table 1 reports our experimental results with synthetic data only, while Table 2 reports the performance of models trained on a mixture of gold and synthetic data. Both tables report results averaged across 5 runs with different data shuffles and model initializations.

The amount of training data for synthetic setups reported in Table 1 is lower than the amount of gold data due to the filtering step being applied to all synthetic sequences (Section 4.2). Specifically, in the 'No *classifier filtering*' setups (*Filter: 'No'* in Table 1), we discard texts for which the output of the model was ill-formatted (i.e., no sequence between inverted commas was in the model output) or sequences were too similar to the original text. The number of training texts further decreases in the 'Classifier filtering' setups (*Filter: 'Yes'* in Table 1), in which we also discard the sequences that did not pass *classifier filtering* (Section 4.2). For these setups, models are on average trained on around two thirds of the amount of data available to the other models, with a different class balance: a large majority of examples that are discarded during this phase are *hateful*, so in the *classifier filtering* setups the synthetic data is composed of very few *hateful* examples. Surprisingly, however, these setups achieve comparable performance with models trained on the original gold data.

Our results show that models trained on synthetic data alone (Table 1) can get close to the performance of classifiers trained on gold data, indicating the potential utility of this approach. However, there is a clear difference between the setups in which *classifier filtering* is employed and those in which it is not. This difference is also visible

		Test data			
		MHS		MDA	HateCheck
		M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		$.811^{\pm.004}$	$.718^{\pm.008}$	$.507^{\pm.027}$	$.386^{\pm.026}$
Gen. Model	Filter				
`Llama-2 Chat 7B` + Gold data	No	$.809^{\pm.005}$	$.719^{\pm.007}$	$.583^{\pm.014}$	$.558^{\pm.029}$
	Yes	$.813^{\pm.003}$	$.723^{\pm.005}$	$.531^{\pm.006}$	$.451^{\pm.010}$
`Mistral 7B Instruct` + Gold data	No	$.812^{\pm.002}$	$.723^{\pm.003}$	$.587^{\pm.009}$	$.609^{\pm.014}$
	Yes	$\mathbf{.816}^{\pm.002}$	$\mathbf{.728}^{\pm.003}$	$.529^{\pm.007}$	$.464^{\pm.011}$
`Mixtral 8x7B Instruct` + Gold data	No	$.811^{\pm.003}$	$.723^{\pm.005}$	$\mathbf{.593}^{\pm.012}$	$\mathbf{.619}^{\pm.010}$
	Yes	$.813^{\pm.003}$	$.723^{\pm.004}$	$.527^{\pm.008}$	$.455^{\pm.025}$

Table 2: Results of `RoBERTa Large` models trained on a mixture of synthetic data + gold data (average of 5 runs ± stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded from the synthetic data. *Filter:Yes* means that *classifier filtering* was applied.

in models trained on a mixture of synthetic and gold data (Table 2), which exhibit similar trends to models trained on synthetic data only, although the differences between setups are less marked. In particular, filtering leads to better performance on the same data distribution (i.e., when testing on the MHS dataset), which could be attributed to the classifier overfitting the original data and misclassifying texts that drift too far from it. Conversely, not filtering typically leads to losses of around .04 F_1 over using actual gold data in in-distribution scenarios, but it can heavily boost performance in out-of-distribution scenarios, with improvements of up to .18 F_1 for the MDA dataset and up to .30 F_1 on HateCheck. This might be due to potential injection of more lexical variety by the LLMs during the paraphrasing process, positively affecting models trained on synthetic data and causing better generalization to out-of-distribution cases.

Better out-of-distribution performance with synthetic data could also be explained by models trained on original gold data and synthetic data potentially learning different types of shortcuts. With the original gold datasets often being constructed using keyword sampling, specific terms are often relied upon by models for classification of hate speech (Ramponi and Tonelli, 2022). On the other hand, with synthetic datasets, models could instead learn different shortcuts (which could potentially be more unpredictable, as we will discuss in Section 6), while becoming more robust to the more 'traditional' shortcuts, which are often a consequence of overfitting the original gold training data. This possible explanation could also account for the slightly lower performance on the same-distribution data when using synthetic training examples, as models trained on synthetic instances would not overfit the original data as much anymore.

6 Intrinsic Evaluation

Our experiments suggest that synthetic data can be useful in making models more robust to out-of-distribution scenarios (cf. Table 1). This would make them advisable for use cases in which hate speech detection has to be performed on target data from a different domain (e.g., genre, topic). However, no in-depth investigation has been carried out so far to highlight what would be the *qualitative* differences between synthetic and gold data for this task. We therefore conduct a qualitative analysis in order to understand what aspects actually play a role in this shift in model performance, to discover what this data contains and, ultimately, if it is truly advisable to use it in real application scenarios.

The qualitative analysis was carried out by two annotators, one male and one female, both with expertise in online language use, hate speech, and LLM-generated text.

The human evaluation focuses on three aspects:

- The *realism* of the synthetic data, i.e., whether a specific message could realistically be found as a social media post;
- To what extent synthetic data creation ensures *hateful content preservation*, i.e., if after paraphrasing the *hateful* messages remain *hateful* (and vice versa for *non hateful* ones);

- Whether the *representation of target identities* is different in the synthetic data compared with the gold data (e.g., if, after paraphrasing, a text that was originally about black women is still about black women, or whether the identity representation was erased).

These aspects can, in fact, have a number of implications on real-world usage of synthetic data for hate speech detection. For instance, if synthetic data is not realistic, it may introduce spurious correlations between certain tokens and labels, making models overfit to lexical items that rarely occur in real-world data (Ramponi and Tonelli, 2022). On the other hand, label preservation is important because the data augmentation process assumes that the label of the original text will be preserved. Indeed, data augmentation gives the opportunity to modify existing data in order to obtain more training samples *without further manual annotation*. However, if a large fraction of the labels changes after augmentation, it might not always be worth it, as classifiers trained on wrongly-labeled synthetic data could have unpredictable performance. Finally, in the frequent cases in which the targets of hate represented in a dataset have been carefully balanced to ensure a fair representation of different groups, changing this distribution through the augmentation process may not be desirable. Moreover, training a classifier on synthetic data in which specific targets of hate have been neglected would potentially affect classifier fairness, hurting already marginalized communities (Xu et al., 2021).

We conduct the human annotation in two steps:

- Annotators are provided with a sample of 500 texts (both gold and synthetic) and asked whether each example appears to be written by a human or an LLM, to estimate how easy it is to spot LLM-written text;

- Annotators are provided with an additional sample of 3,000 synthetic-only examples, i.e., 1,000 texts created by each of the three generative models we employ in our experiments, equally split between the labels. These examples are annotated along a variety of axes, including grammaticality, presence of hate speech, and presence of identity mentions.

The manually annotated data for this work is publicly available at `https://github.com/dhfbk/delving`. Annotation details are reported in the following subsections.

6.1 Realism of Synthetic Texts

The first aspect we investigate is how easy it is to spot synthetic data for a human annotator. While realism is not fundamental for models to recognize hate speech, the ability (or lack thereof) of a human to recognize a text as produced by an LLM might indicate that synthetic texts do exhibit characteristics that cannot fully mimic those of human-written texts. This might, in turn, result in models learning spurious correlations from LLM-written texts, i.e., relying on some expressions or unusual words as shortcuts for classifying posts as hateful.

Human or LLM? In order to assess how real-passing the synthetic texts are, we provide annotators with 500 examples that are a mix of gold texts and texts generated using the three different LLMs that we use in our experiments. To avoid biasing the responses, annotators were not aware of the ratio of real and synthetic examples during the annotation, which is 25% gold and 75% synthetic (i.e., 125 gold examples and 125 synthetic examples for each of the 3 models).

The annotators had an accuracy of 88% in correctly identifying LLM-authored texts, with a precision of 0.83 and a recall of 0.90. The differences across models were small: humans achieved 87%, 90%, and 92% accuracy in correctly identifying synthetic texts generated with `Llama-2 Chat`, `Mistral`, and `Mixtral`, respectively.

Inter-annotator agreement was calculated on 20% of the annotated examples, selected randomly. The annotators agreed 89% of the time, with a Krippendorff's alpha coefficient of 0.73. We believe that the high accuracy might be due to annotators' expertise and familiarity with LLM-generated text. However, this shows that, to an expert eye, synthetic texts might not be quite as realistic as expected. For instance, texts with convoluted constructions and unusual (but polite) lexical choices were often found to be synthetic, such as '*kindly halt this conduct characterized by the blending of unconventional gender identities and feminist ideologies*' (paraphrase of: '*please stop this queer feminist bullsh*t*').

Prompt Failures and Grammar Annotators were asked to label 3,000 synthetic examples (1,000 per model) to report whether *a)* the output did not correctly fulfill the prompt (i.e., the model refused to answer or it answered with a description of the gold text), which we deem a *prompt fail-*

	Llama	Mistral	Mixtral
Prompt failure	14%	11%	5%
Grammar incorrect	1%	2%	1%
World knowledge incorrect	4%	5%	4%

Table 3: Synthetic text realism annotations.

ure, *b)* the grammar was deemed correct / realistic, *c)* the 'world knowledge' exhibited by the model was acceptable. The full guidelines we provided to annotators are reported in Appendix C.

Table 3 reports the percentage of synthetic texts created with each model and annotated according to these three aspects. Overall, there are not large differences across models: all the models produce sequences that are acceptable with regards to grammar and world knowledge in most cases. Prompt failures are more common with `Llama-2 Chat`, while they are much less common with `Mixtral 8x7B`. For prompt failures, the IAA among our annotators was fairly high, with a Krippendorff's alpha of 0.76. While `Llama` is more prone to prompt failures, it might produce texts that appear slightly more realistic to human eyes. This hypothesis is supported by the lower accuracy of humans in identifying `Llama` authored texts compared with the other models, as we have observed.

Tip

Do not assume the synthetic texts will necessarily be human-like, even if they are grammatically correct and plausible, as expert eyes are still able to spot LLM-written text.

6.2 Redistribution of Hateful Texts

The second aspect we investigate is whether models maintain hatefulness during the synthetic data creation process. Ideally, paraphrasing a text classified as hateful should output another text of the same class. We therefore ask annotators to label the same 3,000 synthetic examples following the guidelines for hate speech annotation that were adopted for building the MHS corpus, and then compare the labels with those originally assigned to the gold texts. The difficulty of preserving labels in LLM-based data augmentation has already been attested in the past (e.g., Kumar et al. (2020)), but to our knowledge it has never been qualitatively assessed for subjective tasks such as hate speech detection.

While our aggregation process for the *hate speech* label in the MHS corpus (Section 3) removed the *unclear* label, our annotators could label texts as *hateful*, *non hateful*, and *unclear*, following the original guidelines, reported in Appendix C. For the *hate speech* label, the inter-annotator agreement between our annotators was moderately high, with a Krippendorff's alpha of 0.70.

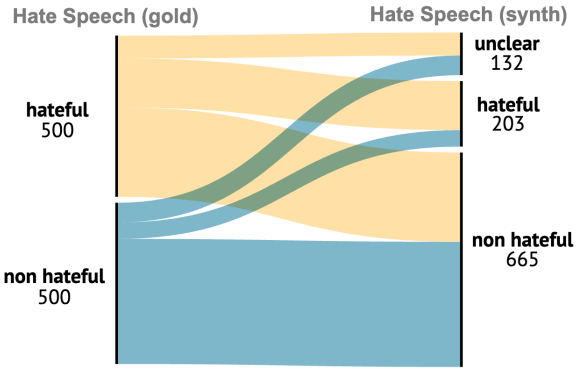

Figure 1: Distribution of hateful and non hateful texts in the manually labeled subset of gold and synthetic data created using the `Mixtral 8x7B Instruct` model.

Overall, tendencies to produce synthetic examples with a different *hate speech* label than their original version are similar across models. For brevity, here we display the statistics for synthetic data produced with `Mixtral 8x7B Instruct`, and refer the reader to Appendix D for `Llama-2 Chat` and `Mistral Instruct 7B`. The overview of the redistribution of labels after the synthetic data creation process is shown in Figure 1.

Across all models, almost half of the examples go through a change of label, with most of these changes regarding texts that are originally *hateful*, which are rendered *non hateful* through the LLM paraphrasing process. We hypothesize this change in label distribution could be at least in part due to the alignment of models, which tends to limit the generation toxic language as they are trained to minimize inappropriate, offensive or unethical uses (Rao et al., 2023). This effect is slightly reduced in the case of Mistral and Mixtral, which do not officially feature any moderation mechanisms compared to Llama 2, which instead officially features safety safeguards. However, the three models all exhibit the same overall tendency to increase the percentage of *non hateful* examples, reducing the overall level of 'hatefulness' present in the dataset. This shows that the presence of safeguards may not be the only factor influencing the ability (or lack thereof) of models to generate offensive content.

Another small portion of examples that go through a change of label in this sense includes *prompt failures*, which are always considered as *non hateful* in our annotation scheme.

Interestingly, there also are a number of examples that transition from being *non hateful* to being *hateful*. In particular, through manually looking at these examples, we note that there are several potential reasons for these changes. Many are cases of clearly sarcastic texts that, through the paraphrasing process, are turned into texts that might sound serious (e.g. *I like that brown people defending their home is 'barbaric'* being turned into *It's savage, in my view, when brown people resist invaders and protect their homes*). Others can be attributed to genuine disagreements between annotators or annotation errors.

Tip

Do not expect paraphrased synthetic texts to maintain the same class distribution as the gold data.

6.3 Redistribution of Target Identities

Given that the representation of different target identities can lead to discrepancies in classification performance across identity groups, risking further marginalization of underrepresented categories (Sap et al., 2019), we also analyze the redistribution of identity categories in the synthetic data. As with the label redistribution experiments, the findings of our analysis generalize across models. Therefore, in this section we only report the statistics for `Mixtral 8x7B Instruct` and refer the reader to Appendix D for the other models.

Annotators are provided the same guidelines as the annotators of the MHS corpus, with 7 categories of identity groups to annotate for both *hateful* and *non hateful* examples: *age*, *disability*, *gender*, *origin*, *race*, *religion*, and *sexuality*. The redistribution of identity group mentions is shown in Figure 2.

The analysis shows that over one third of the examples lose the reference to the original identity group(s) when paraphrased (cf. Figure 2; from any category on the left to *no target* on the right). In particular, the representation of the *gender*, *race*, and *sexuality* categories is heavily reduced, while this reduction is less noticeable for other categories such as *religion* or *disability*. We hypothesize this may also be due to the *alignment* process for these models, which is likely to prevent models from generating hateful messages against the most common targets of hate. Instead, for other categories such as *religion* or *disability*, the model may not have been exposed to them during training, as they are more scarcely represented in widely-used hate speech datasets. Furthermore, creating synthetic paraphrases of texts also appears to reduce the representation of intersectionality, with over half of the gold texts that represent multiple identity categories being either turned into synthetic texts that mention one single identity category or none at all.

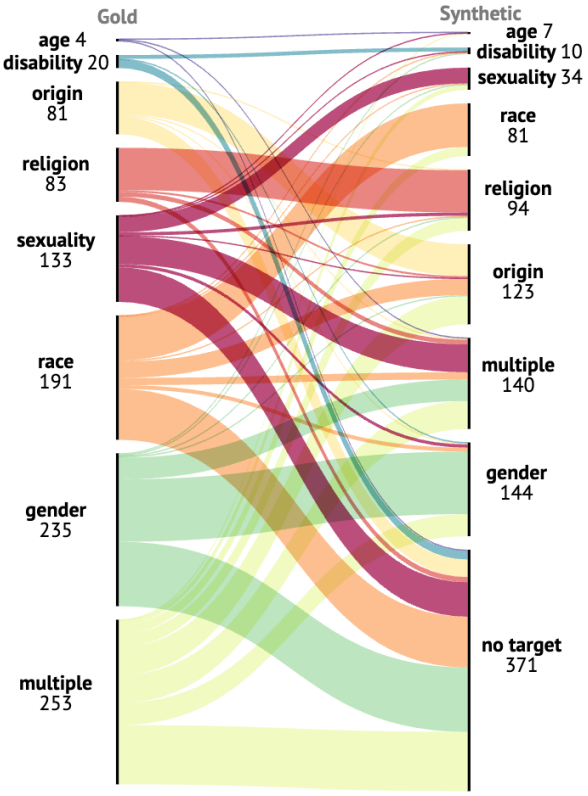

Figure 2: Target identity redistribution with the `Mixtral 8x7B Instruct` model.

Tip

Synthetic data won't necessarily share the same representation of identity groups as the gold data.

To investigate this further, we extract the most informative tokens for the *hateful* class from both the original gold data and the synthetic data with the VARIATIONIST[4] Python library (Ramponi et al., 2024), using the built-in normalized positive weighted relevance metric (`npw_relevance`). Again, given that the findings are similar across all

[4] https://github.com/dhfbk/variationist

Target	Subset	Top-k tokens
AGE	GOLD	f*ck, *ss, b*tch, f*cking, 🖕, sh*t, p*ssy, racist, c*nt, kids
	SYNTH	individuals, individual, woman, children, mother, person, people, sexual, child, women
DISABILITY	GOLD	r*tarded, r*tard, f*cking, f*ck, sh*t, *ss, b*tch, r*tards, people, kill
	SYNTH	individuals, person, foolish, individual, intellectually, impaired, intelligence, mentally, lack, ignorant
GENDER	GOLD	b*tch, f*ck, *ss, f*cking, c*nt, b*tches, sh*t, p*ssy, wh*re, sl*t
	SYNTH	woman, women, person, individuals, individual, promiscuous, ignorant, sex, foolish, sexual
ORIGIN	GOLD	f*ck, f*cking, country, sh*t, people, america, *ss, white, b*tch, american
	SYNTH	individuals, country, people, america, person, individual, return, american, nation, immigrants
RACE	GOLD	n*gga, n*ggas, f*ck, *ss, f*cking, white, sh*t, b*tch, n*gger, 😂
	SYNTH	individuals, people, person, white, individual, racist, black, african, woman, women
RELIGION	GOLD	f*ck, jews, f*cking, sh*t, people, jew, muslim, muslims, white, god
	SYNTH	individuals, people, jewish, individual, jews, muslim, muslims, person, islam, white
SEXUALITY	GOLD	f*ggot, f*ck, f*cking, *ss, f*g, sh*t, f*ggots, gay, b*tch, d*ck
	SYNTH	homosexual, person, individuals, gay, individual, term, behavior, derogatory, effeminate, people

Table 4: Top-k = 10 most informative tokens for the *hateful* class across targets of hate in GOLD and SYNTHETIC posts, calculated using the `npw_relevance` metric of VARIATIONIST. The SYNTHETIC subset refers to texts paraphrased with `Mixtral 8x7B Instruct`.

the generative models, we only report statistics of texts generated using `Mixtral 8x7B Instruct` in Table 4.[5] We report statistics for the other models in Appendix E.

From this analysis, it is clear that LLMs tend to turn any potentially harmful input into its 'safer' counterpart, with all slurs completely disappearing from the list of the most informative tokens for the *hateful* class for each target category. While the synthetic data we analyze actually *is* still useful as training data for classifiers, as we saw in Section 5, it is clear from this analysis that the *content* of this data is largely different from that of the original gold dataset. This might lead to models learning 'shortcuts' for classification, and wrongly assuming that certain commonly used words, such as *woman* or *homosexual*, are to be associated with hateful texts. This can have unpredictable consequences if models trained on synthetic data are actually deployed for the identification of hate speech. We plan to explore this aspect further in future work.

7 Conclusion

In this work, we have carried out an assessment of synthetic data beyond the mainstream classifier performance evaluation, with the goal of linking classifier performance with an intrinsic qualitative analysis. Our aim is to understand the potential risks and drawbacks of using synthetic data for a delicate task such as hate speech detection. While from mere classifier performance synthetic data shows to be helpful in out-of-distribution scenarios, our qualitative analysis proves that we should not take for granted the preservation of key features of gold data in synthetic data. First, synthetic data might introduce spurious correlations due to the language used by models, as it is easily spotted by expert humans. In addition, we showed that the preservation of *hate speech* labels during the augmentation process should not be automatically assumed, even when the data still appears to be useful for training a classifier. Finally, LLM-generated paraphrases of gold data show a drastically different identity category distribution compared with the original data.

Overall, our analysis shows that while classifier performance might show synthetic data to be potentially useful, it can hide potential risks we may often be unaware of.

Limitations

In this work we focus on synthetic data in English and comparatively evaluate generation quality of 3 LLMs. The language choice was mostly driven by the need to analyse classification quality from a cross-dataset perspective and using a target-based angle, which required the availability of specific types of datasets for our experiments. Although we acknowledge that any language model in any language may be affected by the issues that we investigate, the above experimental setting limited our focus to English. Nevertheless, we tried to be cau-

[5] Given the large number of slurs in these lists, we obfuscate profanities according to Nozza and Hovy (2023).

tious in presenting our findings, avoiding overgeneralizations. Furthermore, our manual annotation is carried out by only two annotators, while more annotators could strengthen our findings. Nevertheless, we believe our work to still be potentially useful as a first exploration into the qualitative aspects of synthetic data for hate speech detection.

Impact Statement

The goal of this work is to perform an in-depth analysis of synthetic data for hate speech detection going beyond a simple performance-based evaluation. We therefore try to highlight also the critical risks associated with using this kind of data, which may affect specific targets of hate that are already underrepresented in current datasets. In our study, we use already available datasets and we do not collect, exploit or reshare any personal data. The human annotators involved in the manual evaluation are both affiliated with the authors' institution and performed the task as part of their work activities. This guaranteed a better control over data quality and more awareness of possible annotators' biases. It also provided annotators with a safe environment in which they felt authorized to stop annotating whenever they felt that the task was becoming psychologically taxing.

In general, using LLMs to generate hateful messages is a malicious use of language technologies. In our work, however, we exploit LLMs to ultimately improve hate speech detection systems and to mitigate some issues with existing data and methods. Furthermore, in this paper, we do not propose novel methodologies to generate hateful messages, nor approaches to circumvent model alignment. Also, we do not release the entire generated dataset. Rather, we only make available the set of data which has been manually annotated (3,500 synthetic examples in total)[6] so to provide a test set for future evaluations. This subset does not include the original MHS examples but only their IDs from the original dataset, so that the source MHS corpus should be first retrieved upon approval by its authors to pair the source texts with their synthetic version.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU. This work was also funded by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS).

[6]The manually annotated data is publicly available at https://github.com/dhfbk/delving.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do Not Have Enough Data? Deep Learning to the Rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing Management*, 60(1):103118.

Camilla Casula, Elisa Leonardelli, and Sara Tonelli. 2024. Don't augment, rewrite? assessing abusive language detection with synthetic data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11240–11247, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Camilla Casula and Sara Tonelli. 2023. 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*, pages 3359–3377, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *Preprint*, arXiv:2109.00227.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and

human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Gretel Liz De la Peña Sarracén and Paolo Rosso. 2023. Systematic keyword and bias analyses in hate speech detection. *Information Processing Management*, 60(5):103433.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*, 12:14350–14363.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. 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*, pages 1336–1350, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. 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)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020a. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint*. ArXiv:2009.10277 [cs].

Filip Klubicka and Raquel Fernández. 2018. 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*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A cross-lingual study of homotransphobia on twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24.

Debora Nozza and Dirk Hovy. 2023. The state of profanity obfuscation in natural language processing scientific publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. Association for Computational Linguistics.

Sagi Pendzel, Tomer Wullach, Amir Adler, and Einat Minkov. 2023. Generative ai for hate speech detection: Evaluation and findings. *ArXiv*, abs/2311.09993.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Alan Ramponi, Camilla Casula, and Stefano Menini. 2024. Variationist: Exploring multifaceted variation and bias in written language data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 346–354, Bangkok, Thailand. Association for Computational Linguistics.

Alan Ramponi, Benedetta Testa, Sara Tonelli, and Elisabetta Jezek. 2022. Addressing religious hate online: from taxonomy creation to automated detection. *PeerJ Computer Science*, 8:e1128.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Martin J Riedl, Gina M Masullo, and Kelsey N Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, 107:106262.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. 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)*, pages 41–58, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei and Kai Zou. 2019. 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Cedric Deslandes Whitney and Justin Norman. 2024. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1733–1744, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. 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*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating Online Misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

A Additional Implementation Details

In this section we report additional details about our implementations for reproducibility purposes.

All of the content in gold data we use was anonymized in the form of removing user mentions and URLs, substituting them with '[@USER]' and 'URL', respectively.

We use the HuggingFace library for all model implementations. For the three classifiers, `RoBERTa Large`[7] (355M parameters), `RoBERTa Base`[8] (125M parameters), and `DeBERTa Base`[9] (140M parameters), we use the default hyperparameters of the `TrainingArgs` class, setting batch size to 64, the maximum sequence length to 150, and the learning rate to 5e-6. We train models for 3 epochs. We use all three generative models (`Llama-2 Chat 7B`[10], `Mistral 7B Instruct v0.2`[11], and `Mixtral 8x7B Instruct v0.1`[12]) loaded in 4-bits, and as hyperparameters for generation we use *top-p* decoding = 0.9 and we set the minimum and maximum lengths of the generated sequences to 5 and 300, respectively. The remaining hyperparameters we use are the default ones of the `GenerationConfig` HuggingFace class. Generation took about 300 hours on a Nvidia A40 GPU and training of classifiers took about 20 hours on a Nvidia V100 GPU.

B Additional Results - Extrinsic Analysis

We report the results of the extrinsic analysis with synthetic data only with `RoBERTa base` in Table 5 and those of `DeBERTa base` in Table 6. The classification results of `RoBERTa base` and `DeBERTa base` trained on a mix of synthetic and gold data are instead reported in Tables 7 and 8, respectively.

C Guidelines

In this section we report the annotation guidelines we adopted for the human evaluation.

Hate Speech Below we report the definition of hate speech used for the Measuring Hate Speech Corpus (Kennedy et al., 2020b):

*Does this comment contain hate speech, defined as "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled (e.g., "jew bankers")." The comment contains hate speech if slurs or derogatory language are used toward any person or group regardless of whether or not the target is a member of the slur-target group. (e.g. a comment calling Trump a faggot or a n***** still contains hate speech).*

Possible choices are *yes*, *no*, *unclear*.

Prompt Failure Examples that should be labeled as *prompt failures* fall under two categories:

1. The model was not able to rewrite the message. *Examples:*
   (a) "Sorry, I cannot do this. I am just an AI."
   (b) "This is a hateful message. Let's try being respectful of everyone."
2. Description of original gold, in which the model is not rewriting the text, but is describing it. *Examples:*
   (a) "Here [USER] said something racist."

Grammar This should assess the general morpho-syntactic grammaticality of the message. The register is informal, so attested language use that - although not book-correct - is common in informal scenarios should be considered correct. Code mixing and dialect use should also not be considered ungrammatical. If essential components of the utterance are missing, hindering the understanding of the message, we tend towards no. If they are missing but the message is understandable, we tend towards correctness.

Examples: "y'all, he don't understand" would be correct. "She done did it" would be correct. "the was here" would be incorrect.

World knowledge / plausibility This should assess whether, regardless of morphosyntactic grammaticality, the message makes sense and is realistic/plausible from a semantic standpoint given our knowledge of the world.

Examples: "I will climb a tree with a whiteboard" would not make sense. "I got a paper cut with a steel knife" would not make sense.

Target identity categories If you find a message to be directed at or about a group category, regardless of hate speech presence, select all identity categories that apply.

[7] huggingface.co/FacebookAI/roberta-large
[8] huggingface.co/FacebookAI/roberta-base
[9] huggingface.co/microsoft/deberta-v3-base
[10] huggingface.co/meta-llama/Llama-2-7b-chat-hf
[11] huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[12] huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

			Test data			
			MHS		MDA	HateCheck
		n(train)	M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		30,132	**.805** $^{\pm.003}$	**.708** $^{\pm.006}$	.546 $^{\pm.022}$	.314 $^{\pm.012}$
Gen. Model	Filter					
`Llama-2 Chat 7B`	No	28,289	.742 $^{\pm.004}$	.643 $^{\pm.004}$	.661 $^{\pm.007}$	.490 $^{\pm.016}$
	Yes	21,132	.786 $^{\pm.004}$	.686 $^{\pm.005}$	.595 $^{\pm.012}$	.326 $^{\pm.007}$
`Mistral 7B Instruct`	No	29,344	.743 $^{\pm.007}$	.654 $^{\pm.005}$	.686 $^{\pm.003}$	**.551** $^{\pm.009}$
	Yes	22,453	.784 $^{\pm.005}$	.684 $^{\pm.007}$	.595 $^{\pm.013}$	.337 $^{\pm.009}$
`Mixtral 8x7B Instruct`	No	29,351	.718 $^{\pm.007}$	.632 $^{\pm.006}$	**.696** $^{\pm.005}$	.541 $^{\pm.008}$
	Yes	22,325	.783 $^{\pm.003}$	.687 $^{\pm.004}$	.619 $^{\pm.007}$	.328 $^{\pm.004}$

Table 5: Results of `Roberta Base` models trained on synthetic data only (average of 5 runs ± stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded. *Filter:Yes* means that *classifier filtering* was applied.

			Test data			
			MHS		MDA	HateCheck
		n(train)	M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		30,132	**.809** $^{\pm.002}$	**.717** $^{\pm.005}$	.522 $^{\pm.018}$	.347 $^{\pm.008}$
Gen. Model	Filter					
`Llama-2 Chat 7B`	No	28,289	.736 $^{\pm.004}$	.642 $^{\pm.005}$	.670 $^{\pm.014}$	.597 $^{\pm.019}$
	Yes	21,116	.785 $^{\pm.0066}$	.684 $^{\pm.012}$	.569 $^{\pm.019}$	.332 $^{\pm.016}$
`Mistral 7B Instruct`	No	29,344	.732 $^{\pm.010}$	.643 $^{\pm.007}$	.672 $^{\pm.006}$	.636 $^{\pm.017}$
	Yes	22,445	.782 $^{\pm.005}$	.678 $^{\pm.006}$	.564 $^{\pm.020}$	.387 $^{\pm.008}$
`Mixtral 8x7B Instruct`	No	29,351	.710 $^{\pm.007}$	.626 $^{\pm.004}$	**.697** $^{\pm.007}$	**.638** $^{\pm.014}$
	Yes	22,292	.781 $^{\pm.007}$	.679 $^{\pm.013}$	.579 $^{\pm.028}$	.390 $^{\pm.021}$

Table 6: Results of `DeBERTa Base` models trained on synthetic data only (average of 5 runs ± stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded. *Filter:Yes* means that *classifier filtering* was applied.

Original question for the annotators of the MHS corpus: *Is the comment above directed at or about any individual or groups based on: Race or ethnicity, religion, national origin or citizenship status, gender, sexual orientation, age, disability status.*

D Label and Target Redistribution Across All Models

In this section we report the full results of our human evaluation on label redistribution in the synthetic data across all models. Figures 3 and 5 report the redistribution of hateful content and identity categories, respectively, for the `Llama-2 Chat` model, while figures 4 and 6 report the redistribution of hateful content and identity categories for the `Mistral 7B Instruct` model.

E PMI Analysis

In this section, we report (in Table 9) the most informative tokens for the *hateful* class in synthetic posts created with each of the three models.

		Test data			
		MHS		MDA	HateCheck
		M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		.805 $^{\pm.003}$	.708 $^{\pm.006}$	.546 $^{\pm.022}$	.314 $^{\pm.012}$
Gen. Model	Filter				
`Llama-2 Chat 7B` + Gold data	No	.797$^{\pm.004}$	.702$^{\pm.005}$	.599$^{\pm.003}$	.419$^{\pm.019}$
	Yes	**.806**$^{\pm.003}$	**.714**$^{\pm.005}$	.554$^{\pm.007}$	.330$^{\pm.005}$
`Mistral 7B Instruct` + Gold data	No	.796$^{\pm.001}$	.701$^{\pm.003}$	.596$^{\pm.011}$	.459$^{\pm.010}$
	Yes	**.806**$^{\pm.003}$	**.714**$^{\pm.005}$	.559$^{\pm.008}$	.332$^{\pm.007}$
`Mixtral 8x7B Instruct` + Gold data	No	.794$^{\pm.001}$	.699$^{\pm.002}$	**.608**$^{\pm.004}$	**.470**$^{\pm.012}$
	Yes	**.806**$^{\pm.001}$	**.714**$^{\pm.003}$	.561$^{\pm.011}$	.329$^{\pm.010}$

Table 7: Results of `RoBERTa Base` models trained on a mix of synthetic data + gold data (average of 5 runs ± stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded from the synthetic data. *Filter:Yes* means that *classifier filtering* was applied.

		Test data			
		MHS		MDA	HateCheck
		M-F_1	Hate F_1	M-F_1	M-F_1
Original gold data (MHS)		.809$^{\pm.002}$	.717$^{\pm.005}$	.522$^{\pm.018}$	.347$^{\pm.008}$
Gen. Model	Filter				
`Llama-2 Chat 7B` + Gold data	No	.802$^{\pm.003}$	.712$^{\pm.004}$	**.584**$^{\pm.012}$	.493$^{\pm.017}$
	Yes	.809$^{\pm.002}$	**.720**$^{\pm.004}$	.539$^{\pm.011}$	.349$^{\pm.011}$
`Mistral 7B Instruct` + Gold data	No	.801$^{\pm.002}$	.710$^{\pm.004}$	.570$^{\pm.003}$	.557$^{\pm.002}$
	Yes	**.810**$^{\pm.001}$	**.720**$^{\pm.002}$	.535$^{\pm.007}$	.368$^{\pm.007}$
`Mixtral 8x7B Instruct` + Gold data	No	.797$^{\pm.003}$	.704$^{\pm.003}$	.576$^{\pm.011}$	**.574**$^{\pm.016}$
	Yes	.808$^{\pm.002}$	.718$^{\pm.004}$	.540$^{\pm.007}$	.364$^{\pm.004}$

Table 8: Results of `DeBERTa Base` models trained on a mix of synthetic data + gold data (average of 5 runs ± stdev). Grey cells indicate out-of-distribution performance. *Filter:No* means that only paraphrased sequences too similar to the original ones and ill-formatted texts were discarded from the synthetic data. *Filter:Yes* means that *classifier filtering* was applied.

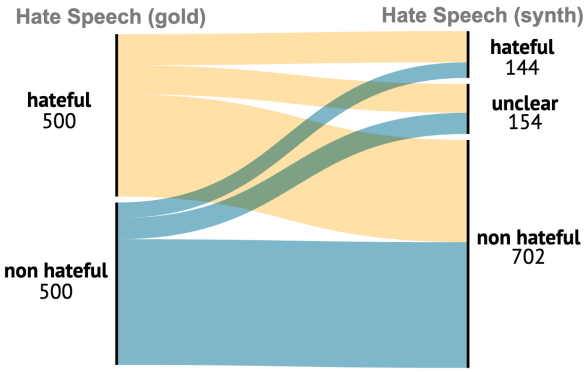

Figure 3: Distribution of hateful and non hateful texts in the subset of gold and synthetic data created using `Llama 2 Chat 7B`.

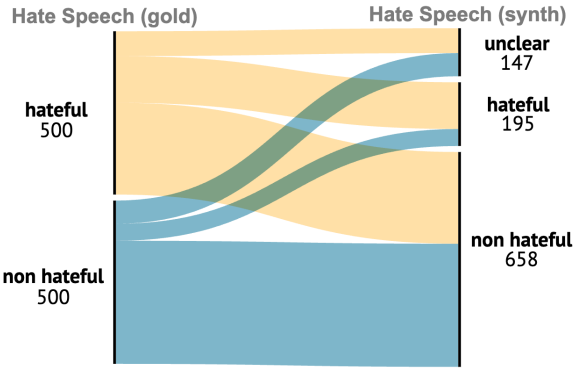

Figure 4: Distribution of hateful and non hateful texts in the subset of gold and synthetic data created using `Mistral 7B Instruct`.

Target	Subset	Top-k tokens
AGE	ORIGINAL	f*ck, *ss, b*tch, f*cking, 🖕, sh*t, p*ssy, racist, c*nt, kids
	LLAMA-2 CHAT 7B	person, language, individuals, people, offensive, individual, sexual, children, mother, life
	MISTRAL 7B	female, woman, children, individuals, anus, person, mother, tiny, outdated, life
	MIXTRAL 8X7B	individuals, individual, woman, children, mother, person, people, sexual, child, women
DISABILITY	GOLD	r*tarded, r*tard, f*cking, f*ck, sh*t, *ss, b*tch, r*tards, people, kill
	LLAMA-2 CHAT 7B	language, person, offensive, individuals, people, derogatory, respectful, disabilities, respect, intellectual
	MISTRAL 7B	person, individuals, individual, woman, foolish, intellectual, girl, intellectually, anonymous, intelligence
	MIXTRAL 8X7B	individuals, person, foolish, individual, intellectually, impaired, intelligence, mentally, lack, ignorant
GENDER	GOLD	b*tch, f*ck, *ss, f*cking, c*nt, b*tches, sh*t, p*ssy, wh*re, sl*t
	LLAMA-2 CHAT 7B	person, language, offensive, sexual, individuals, people, derogatory, respectful, respect, women
	MISTRAL 7B	woman, women, female, person, females, individual, individuals, penis, behavior, foolish
	MIXTRAL 8X7B	woman, women, person, individuals, individual, promiscuous, ignorant, sex, foolish, sexual
ORIGIN	GOLD	f*ck, f*cking, country, sh*t, people, america, *ss, white, american, b*tch
	LLAMA-2 CHAT 7B	individuals, people, country, language, person, derogatory, offensive, america, immigrants, beliefs
	MISTRAL 7B	individuals, america, country, people, return, americans, iran, person, white, american
	MIXTRAL 8X7B	individuals, country, people, america, person, individual, return, american, nation, immigrants
RACE	GOLD	n*gga, n*ggas, f*ck, *ss, f*cking, white, sh*t, b*tch, n*gger, 😂
	LLAMA-2 CHAT 7B	people, individuals, language, person, offensive, derogatory, respectful, respect, race, white
	MISTRAL 7B	individuals, person, people, white, individual, woman, black, racist, behavior, despicable
	MIXTRAL 8X7B	individuals, people, person, white, individual, racist, black, african, woman, women
RELIGION	GOLD	f*ck, jews, f*cking, sh*t, people, muslim, jew, muslims, white, god
	LLAMA-2 CHAT 7B	people, individuals, beliefs, language, offensive, person, respect, including, religion, action
	MISTRAL 7B	individuals, jews, jewish, muslim, person, individual, despicable, muslims, white, islam
	MIXTRAL 8X7B	individuals, people, jewish, individual, jews, muslim, muslims, person, islam, white
SEXUALITY	GOLD	f*ggot, f*ck, f*cking, *ss, f*g, sh*t, f*ggots, gay, b*tch, d*ck
	LLAMA-2 CHAT 7B	language, offensive, sexual, derogatory, person, individuals, people, respect, respectful, lgbtq
	MISTRAL 7B	person, effeminate, homosexual, gay, individual, woman, individuals, penis, derogatory, term
	MIXTRAL 8X7B	homosexual, person, individuals, gay, individual, term, behavior, derogatory, effeminate, people

Table 9: Top-k = 10 most informative tokens for the *hateful* class, according to the PMI metric across targets of hate in GOLD and SYNTHETIC posts paraphrased using `Llama-2 Chat 7B`, `Mistral 7B Instruct`, and `Mixtral 8x7B Instruct`).

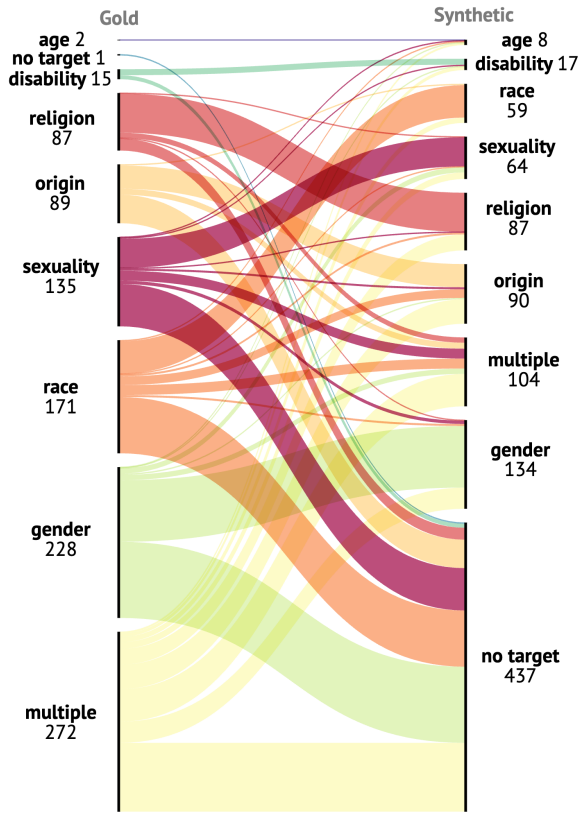

Figure 5: Target identity redistribution in synthetic texts created with `Llama 2 Chat 7B`.

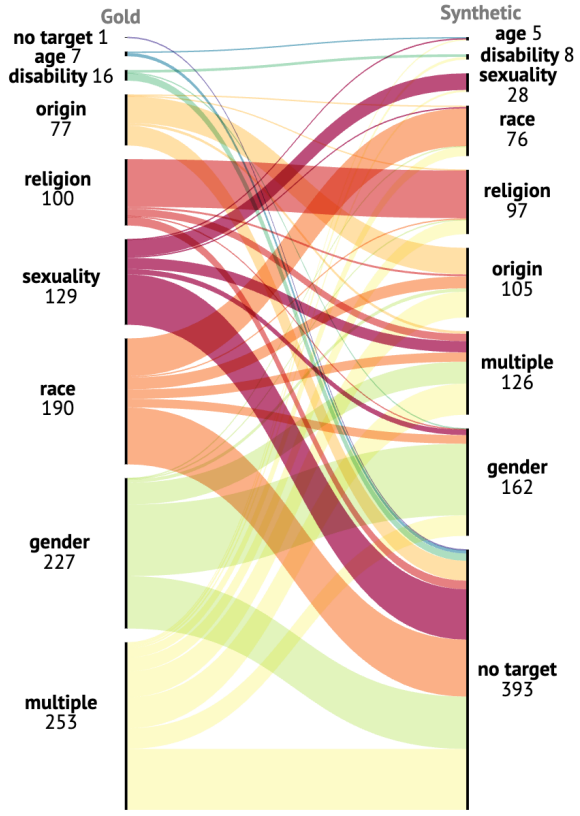

Figure 6: Target identity redistribution in synthetic texts created with `Mistral 7B Instruct`.