LLM-Based Synthetic Datasets: Applications and Limitations in Toxicity Detection

Maximilian Schmidhuber, Udo Kruschwitz
University of Regensburg
maximilian.schmidhuber@stud.uni-regensburg.de, Udo.Kruschwitz@ur.de

Abstract

Large Language Model (LLM)-based Synthetic Data is becoming an increasingly important field of research. One of its promising applications is in training classifiers to detect online toxicity, which is of increasing concern in today's digital landscape. In this work, we assess the feasibility of generative models to create synthetic data for toxic language detection. Our experiments are conducted on six different toxicity datasets, four of whom are hateful and two are toxic in the broader sense. We then employ a classifier trained on the original data for filtering. To explore the potential of this data, we conduct experiments using combinations of original and synthetic data, synthetic oversampling of the minority class, and a comparison of original vs. synthetic-only training. Results indicate that while our generative models offer benefits in certain scenarios, the approach does not improve hateful dataset classification. However, it does boost patronizing and condescending language detection. We find that synthetic data generated by LLMs is a promising avenue of research, but further research is needed to improve the quality of the generated data and develop better filtering methods. Code is available on GitHub; the generated dataset is available on Zenodo.

Keywords: Toxicity, Synthetic Data, Data Augmentation, Large Language Models, Machine Learning

1. Introduction

The rapid advancements in Large Language Models (LLMs), particularly those based on the Transformer architecture (Vaswani et al., 2017), have transformed Natural Language Processing (NLP). These models, trained on massive corpora, demonstrate remarkable generation capabilities to the extent of the fields' leading scientists debating Artificial General Intelligence (Bubeck et al., 2023; Butlin et al., 2023). Efforts to utilize synthetic data are gaining momentum globally. Organizations leverage it to address complex issues such as human trafficking while maintaining data privacy (IOM, 2022)[1]. Synthetic data can also help to alleviate the burden of labelling sensitive datasets (Juuti et al., 2020), has proven valuable in hateful language detection research (Wullach et al., 2021), and has applications in preserving data privacy and bolstering less-resourced NLP tasks (Tennage et al., 2018; Lohr et al., 2018).

This work explores the potential of smaller generational models in data augmentation, specifically to address toxicity detection. We utilize fine-tuned GPT-3 *Curie* instances to generate synthetic text data to enhance downstream ML systems.

Toxicity detection has been a focus of NLP tasks in recent years, in part due to what has been described as a Facebook-fuelled genocide of the Rohingya people in Myanmar (Mozur, 2018). We build upon previous work (Wullach et al., 2021; Meyer et al., 2022b) and investigate the following three research questions:

1. *How effective are classifiers augmented with synthetic data generated by GPT-3 Curie for English hate speech classification, when compared to less-resourced toxicity detection tasks?*

   This explores the variability of synthetic data augmentation effectiveness across tasks and languages. German serves as a less-resourced language contrast, while the subtlety of patronizing language could reveal insights on GPT-3's harm filter and its application in nuanced toxicity detection.

2. *Is it possible to match the performance of classifiers trained on existing toxic language datasets with classifiers exclusively trained on synthetic data?*

   This research question investigates the potential to augment real-world datasets with synthetic ones, which could have implications for privacy and compliance in various fields.

3. *Can synthetic data generated by GPT-3 Curie improve hate speech classifier performance over GPT-2?*

   This research builds on the GPT-2 based methodology of Wullach et al. (2020, 2021). We compare our experimental results on GPT-3 *Curie* generated data to theirs on GPT-2 generated data. We investigate potential improvements due to GPT-3's larger size and capabilities and the potential impact of harm filters on data quality.

[1] https://tinyurl.com/2vs3raf4

TRAC-2024 Workshop, pages 37–51

Our findings indicate that while our generative models offer potential for data augmentation, its hateful language generation capabilities are constrained, likely due to its harm filter. Patronizing non-hateful toxic language detection on the other hand is improved by our methodology. Code [2] is available on GitHub; the generated dataset is available on Zenodo[3].

2. Related Work

2.1. Toxic Language Detection

Toxic language detection is a critical task for mitigating harmful online communication, a focus highlighted by legislation like the EU's Digital Service Act (DSA). According to the DSA, illegal offline conduct is deemed to be illegal online as well, which includes inciting violence or hatred against protected groups based on race, religion or ethnicity. It aims to regulate large (>45m monthly users) social media companies to "protect its users and the users' data".

Toxic language includes interrelated concepts like Hate Speech (Waseem and Hovy, 2016), Abusive Language (Nobata et al., 2016), Cyberbullying (Kumar et al., 2018, 2021), Toxicity (Risch et al., 2021), Misogyny (Kumari and Singh, 2020), or dangerous language (Poletto et al., 2021; Leader Maynard and Benesch, 2016) among others (Fortuna et al., 2020). These definitions can be subjective and often overlap; toxicity and abusiveness are umbrella terms for the distinct, yet related, concepts like Hate Speech (Poletto et al., 2021; Sanguinetti et al., 2018) and Patronizing Language (Pérez-Almendros et al., 2020).

Various research challenges such as SemEval (Basile et al., 2019; Zampieri et al., 2019, 2020; Pavlopoulos et al., 2021), TRAC (Kumar et al., 2018, 2020a), HASOC (Mandl et al., 2019, 2020, 2021) or GermEval (Wiegand et al., 2018; Struß et al., 2019; Risch et al., 2021) address these complexities, emphasizing the need for robust detection methods. The challenge of subjectivity, along with the requirement for large, diverse datasets, motivates the use of data augmentation techniques. LLM-based augmentation approaches offer potential for improving model performance in this domain, as newer models are capable of accurately mimicking human text (Olney, 2023; Mukherjee et al., 2023). However, responsible and ethical use of such techniques is crucial, especially given the potentially harmful nature of toxic language and the biased nature of the models (Zamfirescu-Pereira et al., 2023).

[2] https://github.com/khaliso/thesis
[3] https://zenodo.org/records/10022788

2.2. Data Augmentation

Data Augmentation is defined as the synthesis of new data from existing training data with the objective of improving the performance of a downstream model (Wong et al., 2016). Traditional approaches include mathematical generation (Boedihardjo et al., 2022), synonym replacement (Pappas et al., 2022), and oversampling techniques (Chawla et al., 2002; Maldonado et al., 2019).

In contrast to these traditional approaches, LLM-based data augmentation for specific classification scenarios has the potential to re-define the information theory rule, according to which *processing data can only reduce the amount of information, not add to it* (Beaudry and Renner, 2012). LLMs are trained on vast amounts of data, and their weights and biases incorporate information present in these datasets (Brown et al., 2020). Tasking such a model with replicating a dataset in any way is therefore bound to incorporate parts of this intrinsic knowledge, and can be seen as an abstract knowledge distillation task (Magister et al., 2022).

Applications for synthetic data span code generation (Luo et al., 2023; Gunasekar et al., 2023; Mukherjee et al., 2023), image classification (Krizhevsky et al., 2017; Ramesh et al., 2021; Poole et al., 2022; Betker et al., 2023), robotics (Bousmalis et al., 2023), medicine (Pappas et al., 2022; Ive et al., 2020; Lohr et al., 2018) and toxic language detection (Wullach et al., 2020, 2021; Schmidhuber, 2021; Meyer et al., 2022b; Whitfield, 2021). Various LLMs (e.g., GPT-2 (Anaby-Tavor et al., 2020; Wullach et al., 2020, 2021; Schmidhuber, 2021; Feng et al., 2020; Schick and Schütze, 2021; Whitfield, 2021; Juuti et al., 2020; Papanikolaou and Pierleoni, 2020), GPT-3 (Yoo et al., 2021; Meyer et al., 2022b,a; Shaikh et al., 2022), T5 (Vu et al., 2021) and ChatGPT (Møller et al., 2023)) are suitable for this task, with trade-offs in cost and availability. The currently most widely used models, ChatGPT, are optimized for a chat scenario, while GPT-3 is designed for a more general text completion task. Ye et al. (2023) found that GPT-3 can be as useful for Natural Language Understanding tasks as GPT-3.5, given the wide variety of task designs.

In general, LLM-based data augmentation falls into two main key categories:

1. **Prompt Engineering:** Carefully designed prompts guide LLM output to ensure the generation of relevant, high-quality data. Key considerations include prompt structure, bias mitigation, and evaluation of data variability and coherence (Meyer et al., 2022a; Meister et al., 2023). Additionally, prompt evolution systems can help optimize prompt design (Fernando et al., 2023).

2. **Fine-tuning:** Fine-tuning LLMs on a small, task-specific dataset enables further specialization for data augmentation. This involves potential trade-offs between introducing bias and enhancing the quality of generated data (He et al., 2022; Papanikolaou and Pierleoni, 2020). Fine-tuning can be class-agnostic or class-sensitive.

   (a) **Class-agnostic:** Augmentation focuses on overall data generation, with the class label playing a diminished role. Often, a classifier is used to subsequently assign soft labels (He et al., 2022; Kumar et al., 2020b).

   (b) **Class-sensitive:** LLMs are directly fine-tuned to generate specific class-related data, often requiring further filtering or re-labelling to ensure quality (Yang et al., 2020; Vu et al., 2021).

2.3. Data Augmentation in Toxic Language Detection

In Toxic Language Detection in particular, data augmentation can prove to be a crucial asset for overcoming annotator burden and dataset scarcity (Juuti et al., 2020). GPT-2 has proven effective in this domain (Juuti et al., 2020; Wullach et al., 2020, 2021).

Generalization across toxic language datasets can be limited, as seen in Seemann et al. (2023). This emphasizes the importance of tailoring augmentation to specific datasets. Shaikh et al. (2022) highlight that prompts, if utilized, strongly influence LLM output, with improved instruction following reducing harmful content generation.

Wullach et al. (2020, 2021) offer a foundational methodology for class-specific synthetic data generation with GPT-2. Their filtering with a BERT-based classifier proved effective, and their experiments revealed notable F1 improvements, driven mainly by increased recall while maintaining precision.

Meyer et al. (2022b) built upon their work and used GPT-3 *Curie* for a patronizing and condescending language detection task, achieving improvements over a baseline classifier trained only on original data. Their experiments on unfiltered data highlight the critical role of filtering.

However, there are some gaps in the existing literature. The more recent generative models starting at GPT-3 have only rarely been used for toxic language augmentation, possibly due to cost constraints. Furthermore, there is little recent research focusing exclusively on synthetic data. This approach emphasizes preservation over performance gains, and could lead to improvements in data availability, privacy preservation and compliance.

2.4. Ethical Considerations

The ethical considerations in the deployment of LLM-based data augmentation are vast. Utilizing an LLM to generate synthetic data gives the LLM immense leverage over the task at the end of the pipeline. It is therefore paramount to be well-informed over any biases, tendencies, and privacy concerns the LLM might pose.

1. **Privacy:** While synthetic data aims to mitigate privacy breaches, there is no guarantee for superior performance over traditional methods. Researchers must critically assess the privacy-utility trade-off. Additionally, LLMs trained on private data can potentially leak that data when prompted (Perez et al., 2022).

2. **Toxicity & Hate:** Generating toxic content can aid in its detection, but also poses risks for misuse. Safeguards against creating harmful AI tools are crucial. Red-teaming for instance is an active research area aiming to identify LLM vulnerabilities (Perez et al., 2022; Ganguli et al., 2022). Mitigating toxic tendencies in LLMs themselves remains an open problem (Gehman et al., 2020).

3. **Time:** While language changes slowly, it also changes constantly (Aitchison, 2005). Especially in toxic language detection, what is considered hurtful or patronizing is susceptible to change, e.g. the statement "She is a bossy woman" carries a slightly different connotation than "He is a bossy man" today, but might not in the future. If a data point was attributed a certain label some time ago, it might no longer be true today.

4. **Model Bias:** LLMs inherit biases from training data, affecting both generated data and subsequent classifiers (Nangia et al., 2020; Blodgett et al., 2020; Abid et al., 2021; Bommasani et al., 2022). Bias detection and mitigation techniques are essential. Sycophancy and deceptive reasoning of LLMs further complicate the issue (Turpin et al., 2023; Nanda et al., 2023).

5. **Democratization of AI:** Synthetic data could break reliance on proprietary datasets, making AI research more accessible. However, if biased LLMs create synthetic data, this will amplify issues rather than actually addressing them. (Paullada et al., 2021; Solaiman and Dennison, 2021).

3. Methodology

This research employs GPT-3 *Curie* for synthetic data generation, building upon the works of Wullach

et al. (2020, 2021) and Meyer et al. (2022b), while adapting them to the task at hand.

3.1. Datasets

We evaluated six datasets. Davidson (Davidson et al., 2017), Founta (Founta et al., 2018), HatEval (Basile et al., 2019) and Stormfront (de Gibert et al., 2018) are also investigated by Wullach et al. (2020, 2021) and focus on English Hate Speech detection. The GermEval dataset (Risch et al., 2021) adds German Toxic Language detection, while the PCL dataset (Pérez-Almendros et al., 2020) tackles subtle patronizing and condescending language. This selection allows both a comparison to prior experiments and explores LLM performance on different Toxic Language variations.

3.2. Classifiers

RoBERTa (Liu et al., 2019), AlBERT (Lan et al., 2019), HateBert (Caselli et al., 2021a) BERT and multilingual BERT (Devlin et al., 2018) were the classifiers evaluated for their performance on both full and undersampled original training sets. HateBert is a BERT model fine-tuned on English hateful Reddit comments.

3.3. Generative Model

We selected GPT-3 *Curie* (Brown et al., 2020) as our Generative Model. While GPT-3 *DaVinci* was the strongest available model that could be fine-tuned at the time of experimentation, GPT-3 *Curie* offers comparable performance while being both a lot more economically feasible and building upon previous work with PCL data (Meyer et al., 2022b). Open-source alternatives like GPT-J Wang and Komatsuzaki (2021) or GPT-NeoX-20B (Black et al., 2022) were considered, but were either less powerful or more computationally demanding.

3.4. Pre-processing

During pre-processing, all datasets were transformed to be binary (0: non-toxic, 1: toxic). Afterwards, the data D_{orig} was split into 80/20 train-test sets $D_{orig\text{-}train}$ and $D_{orig\text{-}test}$ where no testing data was supplied, preserving class imbalance. We also created undersampled training sets $D_{orig\text{-}us}$. All datasets were shuffled for unbiased validation.

3.5. Data Generation

The data generation pipeline was inspired by Wullach et al. (2020, 2021). $D_{orig\text{-}train}$ was split by class label. This split results in two datasets, $D_{orig\text{-}0}$ and $D_{orig\text{-}1}$, to fine-tune two GPT-3 *Curie* models, respectively. The OpenAI API expects a .jsonl document in the format of prompt-completion pairs. In the next step, we therefore transform both datasets to fit this schema. In accordance with the pipeline proposed by Wullach et al. (2020), we used an empty (") prompt. For the completion section, the text samples from the datasets were used.

These datasets are then used to fine-tune a GPT-3 *Curie* model via the OpenAI API, resulting in $FT_{orig\text{-}0}$ and $FT_{orig\text{-}1}$. The fine-tuned models are prompted (") to generate a total of 40.000 synthetic samples per class-label, resulting in $D_{synth\text{-}0}$ and $D_{synth\text{-}1}$. The maximum token length of each generated output was set to the average token length of the corresponding $D_{orig\text{-}0}$ and $D_{orig\text{-}1}$. We furthermore removed any tabs the models had created, as the samples $D_{synth\text{-}0}$ and $D_{synth\text{-}1}$ were saved in a .tsv file, and replaced them with a space (' '). For the synthetic PCL datasets, only the missing synthetic data to get to 40,000 raw synthetic samples per label was generated, as we had access to the synthetic data created by Meyer et al. (2022b). The total cost of synthetic data generation was $269,80 USD.

3.6. Filtering

Filtering is crucial for ensuring the quality of class-conditioned synthetic data, as noted by Meyer et al. (2022b); Wullach et al. (2020, 2021) and Anaby-Tavor et al. (2020). Our filtering method slightly differs from Wullach et al. (2020, 2021). Instead of a BERT model, we fine-tuned all five evaluated classifiers on $D_{orig\text{-}train}$ and evaluated them on $D_{orig\text{-}test}$. We then used the strongest performing baseline classifier to filter the corresponding $D_{synth\text{-}0}$ and $D_{synth\text{-}1}$. Samples mismatching their intended label (e.g., label 1 data generated by $FT_{orig\text{-}0}$) or with confidence scores below 0.7 were discarded, following Wullach et al. (2021). These samples were then combined to form D_{synth}.

While our initial goal was to have 40,000 cleaned synthetic samples per dataset, filtering loss varied greatly. As can be seen in Table 1, up to 96% of data was discarded. Compared to earlier work (Meyer et al., 2022b; Wullach et al., 2020), our $FT_{orig\text{-}1}$ model generated a lot less toxic data.

3.7. Experiments

Due to this high rejection rate, reaching 40,000 samples for all datasets was not economically feasible. To maximize the use of the available synthetic data, we designed three experiments that were conducted using the best baseline classifier: fine-tuning on all available data, only on synthetic data, and synthetic oversampling. To check for robustness, the runner-up classifier from the baseline selection process was also evaluated on the Composite experiments. Significance testing was done

Dataset	Synthetic 0	Synthetic 1	Synthetic filtered 0	Synthetic filtered 1
Davidson (Davidson et al., 2017)	43479	42540	41790	1521
Founta (Founta et al., 2018)	40996	41269	40782	5268
HatEval (Basile et al., 2019)	43758	41273	40991	22587
Stormfront (de Gibert et al., 2018)	43536	40259	41523	22988
GermEval (Risch et al., 2021)	40334	40935	34801	5154
PCL (Pérez-Almendros et al., 2020)	44073	44642	42919	10474

Table 1: Number of synthetic samples before and after filtering

through cross-validation using Bonferri-corrected paired t-tests.[4]

3.7.1. Composite (C)

Evaluates whether adding D_{synth} to the $D_{orig\text{-}train}$ improves classifier performance.

Here, the classifier is either fine-tuned on D_{synth} along with $D_{orig\text{-}train}$, or uses an undersampled version (US) of both, $D_{orig\text{-}us}$ and $D_{synth\text{-}us}$. The evaluation is conducted on $D_{orig\text{-}test}$.

3.7.2. Synthetic (S)

The classifier is only fine-tuned on D_{synth} or $D_{synth\text{-}us}$. Here, we also implemented 5-fold cross-validation for statistical testing, which was conducted on $D_{orig\text{-}train}$. The evaluation is conducted on $D_{orig\text{-}test}$.

3.7.3. SMOTE-like

Inspired by previous work (Chawla et al., 2002; Meyer et al., 2022b; Maldonado et al., 2019), we use D_{synth} to balance a skewed $D_{orig\text{-}train}$ before fine-tuning. This method uses synthetic samples to balance the minority class, as displayed in Pseudocode 1. The evaluation is conducted on $D_{orig\text{-}test}$.

Algorithm 1 Adjust Dataset Lengths

$D_{\text{comp-1}} = D_{\text{orig-1}} + D_{\text{synth-1}}$
if len($D_{\text{comp-1}}$) < len($D_{\text{orig-0}}$) **then**
    $D_{\text{orig-0}} = D_{\text{orig-0}}[: \text{len}(D_{\text{comp-1}})]$
else if len($D_{\text{comp-1}}$) > len($D_{\text{orig-0}}$) **then**
    $D_{\text{synth-1}} = D_{\text{synth-1}}[: \text{len}(D_{\text{orig-0}}) - \text{len}(D_{\text{orig-1}})]$
    $D_{\text{comp-1}} = D_{\text{orig-1}} + D_{\text{synth-1}}$
end if

4. Evaluation and Results

4.1. Baseline Classifier Selection

Surprisingly, most of our baseline classifiers achieved higher macro F1, Precision and Recall than those reported by Wullach et al. (2021), Meyer et al. (2022b) and Schmidhuber (2021). Only the HatEval classifiers consistently returned lower performance.

HateBert emerged as the most consistently strong performer, being either the best or runner-up across all datasets. This suggests that 'hateful' embeddings are effective for toxic language detection, even transcending language barriers in the case of GermEval. AlBERT, however, fell behind expectations, as it never achieved a top or runner-up position.

While there is no clear correlation between dataset size, imbalance and whether the full training set or undersampled training data is optimal, undersampled classifiers often yielded higher recall. This is important, as minimizing false negatives in toxic language detection is critical.

4.2. Composite (C)

In the case of the Hate Speech datasets, the Composite approach generally yielded results between those of classifiers trained on $D_{orig\text{-}train}$ and its undersampled counterpart trained on $D_{orig\text{-}us}$, with a few classifiers performing a lot worse. This pattern was observed across Founta, Stormfront, Davidson, and HatEval. Pre-processing errors (e.g., HatEval D_{synth} containing Spanish samples not used by Wullach et al. (2020)) may have affected performance.

For the Toxic datasets, mBert trained on $D_{orig\text{-}us}$ performed best for GermEval. Issues with GPT-3 *Curie* generating non-English text are hinted at by substantial filtering of label 0 data. However, HateBert fine-tuned on the undersampled data performed well, even though the presence of synthetic data appears to be a hindrance in this case. Only the experiments on the PCL dataset (patronizing and condescending language) showed modest F1 score improvements. This suggests GPT-3's capability to provide meaningful variations for this subtle form of toxicity, possibly due to the harm filter being less restrictive for non-hateful content.

GPT-3 *Curie* generated Synthetic data appears to have limited benefit for heavily imbalanced hate speech datasets. This could point to GPT-3's harm filter limiting the generation of novel harmful content. Also, pre-processing errors in some datasets likely impacted the results. We will provide more details in the Limitations section. Overfitting may explain some cases where only one label was pre-

[4] Detailed settings and results can be found in the project repository.

Table 2: Recall and Macro F1 for full and undersampled Composite and SMOTE experiments in comparison to the original. The highest result per dataset is marked **bold**, and the runner-up ***bold and italic***

Dataset	Classifier	Original		O. US		Composite		C. US		SMOTE	
		R	F1	R	F1	R	F1	R	F1	R	F1
Founta	BERT	78.08	**81.87**	86.31	70.69	80.54	80.72	83.68	67.12	85.20	71.98
	HateBert	76.30	***80.97***	85.76	69.23	79.52	79.63	84.78	70.58	85.33	71.54
Stormfront	HateBert	72.87	71.16	83.68	**83.63**	67.99	65.80	81.38	81.32	78.87	78.87
	RoBERTa	0.5	33.3	82.01	***81.96***	49.79	35.98	51.26	41.99	55.02	44.39
Davidson	HateBert	92.89	**92.64**	91.38	87.78	81.82	84.15	90.52	87.70	91.78	89.62
	BERT	90.79	***90.93***	90.53	87.87	55.63	54.06	89.43	85.46	89.65	89.11
HatEval	HateBert	56.32	43.44	59.14	**49.02**	56.02	43.23	50.70	36.10	57.99	46.91
	RoBERTa	58.82	***48.72***	55.69	42.16	50.0	36.71	50.0	36.71	54.48	40.48
GermEval	mBert	70.92	70.81	81.26	**81.25**	50.0	34.3	58.03	54.96	67.42	67.13
	HateBert	51.51	38.73	79.17	***78.98***	60.53	60.17	76.29	76.01	65.80	65.69
PCL	Bert	69.39	71.78	81.23	65.94	50.0	47.51	82.84	64.2	71.61	***73.10***
	HateBert	68.77	71.51	79.5	64.74	69.01	72.28	80.04	59.40	74.96	**73.72**

Table 3: Mean score (Standard Deviation)—in percent, for original and synthetic classifiers, calculated on original validation sets in 5-fold cross-validation. Significantly **worse** F1 scores of synthetic classifiers compared to their original counterparts are marked **bold**.

	Original				Synthetic			
Dataset	A	P	R	F1	A	P	R	F1
Founta								
Bert	93.27 (1.9)	70.24 (22.5)	64.28 (13.9)	65.93 (17.1)	91.70 (0.4)	73.64 (0.6)	79.10 (2.6)	75.94 (1.3)
Bert US	85.15 (0.6)	85.14 (0.6)	85.13 (0.6)	85.13 (0.6)	83.27 (2.1)	84.23 (1.9)	83.24 (2.1)	83.13 (2.2)
Stormfront								
HateBert	92.34 (1.2)	73.12 (15.8)	69.07 (10.9)	70.48 (12.9)	68.85 (12.2)	54.17 (5.1)	64.74 (8.3)	50.68 (2.2)
HateBert US	84.49 (2.7)	84.75 (2.5)	84.50 (2.6)	84.44 (2.8)	54.39 (8.7)	35.41 (22.7)	53.71 (8.3)	**40.25 (14.9)**
Davidson								
HateBert	94.21 (0.5)	92.24 (0.5)	92.64 (0.9)	92.42 (0.6)	68.28 (3.5)	45.00 (4.9)	48.55 (1.5)	**45.72 (2.3)**
HateBert US	92.57 (1.0)	92.61 (0.9)	92.56 (1.0)	92.56 (1.0)	76.53 (3.2)	80.45 (2.3)	76.53 (3.1)	**75.69 (3.5)**
HatEval								
HateBert	68.28 (13.85)	69.41 (11.01)	73.36 (7.3)	65.84 (15.33)	81.47 (6.8)	77.07 (6.7)	82.82 (5.0)	77.81 (6.7)
HateBert US	82.03 (0.1)	81.84 (0.5)	82.46 (0.9)	82.13 (0.2)	82.57 (0.9)	83.10 (0.9)	82.56 (0.9)	82.49 (0.9)
GermEval								
mBert	60.62 (6.5)	49.48 (20.0)	58.23 (8.6)	52.32 (15.68)	56.63 (0.5)	28.32 (0.3)	50.0 (0)	36.16 (0.2)
mBert US	60.42 (5.5)	55.33 (16.8)	60.20 (5.9)	56.90 (13.0)	62.49 (4.9)	64.83 (4.5)	62.57 (4.2)	60.95 (5.0)
PCL								
Bert	90.66 (0.5)	66.94 (12.2)	62.78 (7.3)	64.20 (9.4)	78.39 (6.9)	61.85 (1.9)	74.78 (1.5)	62.74 (4.2)
Bert US	81.55 (1.9)	81.47 (1.9)	81.49 (1.9)	81.45 (1.9)	74.43 (1.9)	78.30 (1.2)	74.45 (0.9)	**73.44 (1.6)**

dicted (R=50.0), particularly in imbalanced training scenarios.

4.3. Synthetic (S)

We trained the base version of the winning baseline classifier of each dataset on $D_{\textit{orig-train}}$, $D_{\textit{orig-us}}$, $D_{\textit{synth}}$ and $D_{\textit{synth-us}}$ in 5-fold cross-validation. The models fine-tuned on synthetic data were validated on the corresponding original dataset. In Table 3, we give an overview of the cross-validation results. When applying paired t-tests to macro F1 results with $p < 0.0042$ [5] we get four significant results for 3 different datasets, all of which mark a significant performance decrease.

As can be seen in Table 3, synthetic-only macro F1 for Stormfront was significantly worse for $HateBert_{\textit{synth-us}}$ (t(4) = 6,51, p = 0.0029) when compared to $HateBert_{\textit{orig-us}}$, while the difference between $HateBert_{\textit{synth}}$ and $HateBert_{\textit{orig-train}}$ was found to be not significant (t(4) = 3,94, p = 0.0170).

For Davidson, macro F1 of $HateBert_{\textit{orig-train}}$ was significantly higher than that of $HateBert_{\textit{synth}}$ (t(4) = 46,09, p < .001), and $HateBert_{\textit{orig-us}}$ outperformed $HateBert_{\textit{synth-us}}$ (t(4) = 12.12, p < .001).

In the case of PCL, the model trained on $D_{\textit{synth}}$ did not significantly lag behind its original counterpart, while macro F1 of $Bert_{\textit{orig-us}}$ was significantly higher than $Bert_{\textit{synth-us}}$ (t(4) = 9.62, p < .001).

[5] To account for multiple comparisons, we applied a Bonferroni-correction of $p = 0.05/12 = 0.0042$ to set the threshold for significant results.

In the cases of the Founta, HatEval and GermEval datasets, however, the models trained on the synthetic data variations did not significantly lag behind their original counterparts.

4.4. SMOTE-like

The SMOTE approach consistently performed well across all tested datasets, being the top-performing or runner-up approach for the synthetic data experiments in HatEval, Davidson, Stormfront, PCL and GermEval. Most notably, HateBert fine-tuned on the SMOTE-like dataset achieved the highest result on any experiment on PCL data, achieving a higher F1 score than the classifiers trained on original data.

4.5. GPT-3 vs. GPT-2

As displayed in Table 4, we find that our baseline models are surprisingly strong. We achieved higher macro F1 scores than previous work (Wullach et al., 2020, 2021; Meyer et al., 2022b) in three of the four datasets using either the full or undersampled training set. Our experiments involving synthetic data on the other hand, returned mixed results. The macro F1 of Davidson $D_{comp\text{-}us}$ is comparable to that reported by Wullach et al. (2021), and Founta $D_{comp\text{-}train}$ exceeded all classification results reported by them on this dataset. On the other hand, the experiments involving RoBERTa saw a steep decline in performance. We also need to note that Wullach et al. (2021) achieved stronger macro F1 results on both our baseline and composite experiments on the HatEval dataset, while Precision and Recall are similar.

HateBERT emerged as the best or second-best classifier on all datasets, even on the German GermEval set. This underscores the power of biasing models towards hate speech, even when the model is trained in a language it is not evaluated on. We find no clear pattern for undersampling. The benefits in F1 score of undersampled vs. full datasets vary across datasets, with no clear link to dataset size or imbalance. Undersampled classifiers do, however, often show higher recall, making them ideal if false negatives are of high concern.

GPT-3 *Curie* generated synthetic data appeared to have a detrimental impact on some, but not all, classifier performances.

5. Discussion

While our works build on Wullach et al. (2021) and Meyer et al. (2022b), there are a few key differences. We utilize undersampling and SMOTE-like techniques, and investigate synthetic-only training scenarios. Let us revisit our research questions:

1. *Are classifiers augmented with synthetic data generated by GPT-3 Curie for English hate speech classification more effective, when compared to less-resourced toxicity detection tasks?*

   English hate speech classifiers saw performance decreases with synthetic data. For German toxic language, multilingual BERT performed best at baseline, but HateBert outperformed it on synthetic data. This suggests possible cross-linguistic hate speech pattern recognition. The best results were seen on the subtle patronizing and condescending language (PCL) dataset, especially on synthetic oversampling.

   Conclusion: H1 is partially accepted. The impact of GPT-3 *Curie* generated synthetic data varies across tasks and languages.

2. *Is it possible to match the performance of classifiers trained on existing toxic language datasets with classifiers exclusively trained on synthetic data?*

   Synthetic-only classifiers underperformed significantly on Davidson and the undersampled PCL and Stormfront datasets. No significant impact was seen on the remaining datasets.

   Conclusion: H2 is partially rejected, as the results were dataset-dependent. A possible explanation is GPT-3's harm filter, which would limit the generation of novel harmful content, making the approach less effective for explicitly hateful datasets.

3. *Can synthetic data generated by GPT-3 Curie improve hate speech classifier performance over GPT-2?*

   GPT-3 *Curie* generated data negatively impacted English hate speech classifier performance compared to baseline classifiers. This contrasts with the findings of Wullach et al. (2021) using GPT-2 generated data. This negative impact could be explained by either the harm filter of GPT-3 *Curie* or by our stronger baselines.

   Conclusion: H3 is rejected. GPT-3 *Curie*, following our methodology, does not achieve stronger performance than GPT-2 for English hate speech classifier performance.

We also find that the data preparation approach made as much, if not more, difference than synthetic data. The SMOTE-like approach consistently performed well, and training models on both the full training data and undersampled training data had a positive impact in our experiments. If one approach had failed due to under- or overfitting, the other often delivered a usable model. Finally,

Table 4: Comparison to Wullach et al. at Base and Gen:80K

Dataset	Classifier	Metric	Original		Composite	
			Wullach et al.	Own results	Wullach et al.	Own results
Founta	Bert	P	73.0	66.85 (O. US) / 87.27 (O)	84.9	64.38 (C. US) / 80.91 (C)
		R	65.0	86.31 (O. US) / 78.07 (O)	67.8	83.68 (C. US) / 80.54 (C)
		F1	68.8	70.69 (O. US) / 81.87 (O)	75.4	67.12 (C. US) / 80.72 (C)
Stormfront	Bert	P	60.9	70.73 (O. US) / 74.8 (O)	-	-
		R	56.2	70.71 (O. US) / 57.95 (O)	-	-
		F1	58.5	70.70 (O. US) / 49.35 (O)	-	-
	RoBERTa	P	80.9	82.22 (O. US) / 25.0 (O)	87.2	53.47 (C. US) / 48.48 (C)
		R	63.7	82.01 (O. US) / 50.0 (O)	73.6	51.26 (C. US) / 49.79 (C)
		F1	71.3	81.96 (O. US) / 33.33 (O)	79.8	41.99 (C. US) / 35.98 (C)
Davidson	Bert	P	98.1	86.10 (O. US) / 91.07 (O)	87.5	83.45 (C. US) / 74.62 (C)
		R	70.6	90.53 (O. US) / 90.79 (O)	86.8	89.43 (C. US) / 55.63 (C)
		F1	82.1	87.87 (O. US) / 90.93 (O)	87.1	85.46 (C. US) / 54.06 (C)
HatEval	Bert	P	69.6	66.78 (O. US) / 68.27 (O)	-	-
		R	53.5	55.90 (O. US) / 56.2 (O)	-	-
		F1	60.5	43.26 (O. US) / 43.37 (O)	-	-
	RoBERTa	P	64.0	68.77 (O. US) / 68.06 (O)	70.6	29.00
		R	64.2	55.69 (O. US) / 58.82 (O)	80.8	50.0
		F1	64.1	42.16 (O. US) / 39.12 (O)	75.4	36.71

HateBert performed well on all challenges related to toxicity detection, regardless of language or the complexity of the task it was tested it on; its use-case can therefore possibly be extended beyond hate to the field of toxicity detection in general.

6. Conclusion and Future Work

This research demonstrates the potential and limitations of GPT-3 *Curie* for synthetic toxic data generation. We find that strict filtering is crucial, and performance may still be lower than using original data alone. GPT-3 *Curie* is feasible with non-hateful toxic language, providing a potential avenue of research when original data is limited. We further note the importance of utilizing both full and undersampled versions of a dataset, and underline the power of synthetically oversampling the minority class (SMOTE) for stability.

There is a plethora of research avenues for future work. Our experiments listed in Tables 2 and 4 need to be cross-validated and tested for significance. ANOVA could be utilized to test for significance in the relationships between using the full datasets, undersampling, and the SMOTE-like approach. An exploratory data analysis using methods like unique word comparison, ROGUE-L and cosine similarity to investigate the discrepancy in results between and within the original and synthetic datasets is recommended. Filtering techniques beyond our approach could be tested and compared, including more traditional machine learning concepts like XGBoost or Naive Bayes.

We find GPT-3 *Curie* to be not suitable to generate synthetic hateful language, likely due to its harm filter. However, other generative models, both proprietary and open-source, could be fruitful. Alternative generation techniques, such as using soft labels (Yang et al., 2020; He et al., 2022) or class-agnostic approaches based on prompting or fine-tuning, offer a more resource-friendly path and could be investigated. Crucially, a thorough evaluation of our approach using privacy-preservation metrics is needed to assess feasibility.

All things considered, LLM-based data augmentation is an immensely powerful tool that promises to remove some of the barriers in the way of science. Before we get there, however, there is still some work to be done, and this paper is hopefully a step in this direction. We need to thoroughly understand model biases and potential pitfalls through rigorous tests like red-teaming (Perez et al., 2022; Ganguli et al., 2022). We need to understand a model structure for it to be as effective as possible, i.e. we find it is not recommended to generate harmful data with a model that has a harm filter with no accessible way of circumventing it for research.

7. Limitations

The ethical considerations outlined in the ethics section must be reiterated. Model biases can potentially be amplified in our pipeline, where a potentially biased model generates synthetic data, filtered by another biased model, only to train yet another biased classifier.

Our generative model may have been trained on some of the evaluated datasets (except for PCL and GermEval datasets, which were published after GPT-3's knowledge cutoff), impacting the evaluation of synthetic data.

The current binary classification approach presents scalability issues for multi-label datasets. Alternative generation methods that are class-

agnostic or use a one-model approach should be explored to address this limitation.

Our study also faced several limitations that warrant acknowledgement. An error led to overlaps between training and test data for the GermEval (75/609 test cases) and Founta (163/11764 test cases) data entries. This contamination, especially pronounced in GermEval, may affect the validity of the results. The HatEval datasets used to fine-tune GPT-3 *Curie* included Spanish data due to a preprocessing error, which hinders direct comparisons with prior work. No Spanish data was contained in later steps of the experiments. And finally, as seen in Table 4, we did not conduct all experiments on Bert, AlBERT and RoBERTa that were done by Wullach et al. (2021) due to time constraints.

8. Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback.

9. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models.

Jean Aitchison. 2005. Language change. In *The Routledge Companion to Semiotics and Linguistics*, pages 111–120. Routledge.

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63.

Normand J. Beaudry and Renato Renner. 2012. An intuitive proof of the data processing inequality.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. 2022. Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data. *Foundations of Computational Mathematics*, pages 1–48.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. 2023. Robocat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. 2023. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2021a. Hatebert: Retraining bert for abusive language detection in english. *WOAH 2021*, page 17.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021b. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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

Kelly Dekker and Rob van der Goot. 2020. Synthetic data for english lexical normalization: How close can we get to manually annotated data? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6300–6309.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. *arXiv preprint arXiv:2010.01794*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Harry G Frankfurt. 2005. *On bullshit*. Princeton University Press.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny

Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Edward S Herman and Noam Chomsky. 1988. *Manufacturing consent: The political economy of the mass media*. Vintage.

Microsoft Research IOM. 2022. IOM and Microsoft release first-ever differentially private synthetic dataset to counter human trafficking. https://www.microsoft.com/en-us/research/blog/.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. *arXiv preprint arXiv:2009.12344*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018. Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 58–65.

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021. The comma dataset v0. 2: Annotating aggression and bias in multilingual social media discourse. *arXiv preprint arXiv:2111.10390*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020a. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020b. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Kirti Kumari and Jyoti Prakash Singh. 2020. AI_ML_NIT_Patna@ TRAC-2: Deep learning approach for multi-lingual aggression identification. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 113–119.

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

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, 9(3).

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*.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution—a case study to get around iprs and privacy constraints featuring the german jsyncc corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Sebastián Maldonado, Julio López, and Carla Vairetti. 2019. An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 29–32.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Sanguinetti Manuela, Comandini Gloria, Elisa Di Nuovo, Simona Frenda, Marco Antonio Stranisci, Cristina Bosco, Caselli Tommaso, Viviana Patti, Russo Irene, et al. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling.

Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022a. Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6.

Selina Meyer, Maximilian Schmidhuber, and Udo Kruschwitz. 2022b. Ms@ iw at semeval-2022 task 4: Patronising and condescending language detection with synthetically generated data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 363–368.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Paul Mozur. 2018. A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Andrew M Olney. 2023. Generating multiple choice questions from a textbook: Llms match human performance on most metrics.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. 2022. Data augmentation for biomedical factoid question answering. *arXiv preprint arXiv:2204.04711*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Carla Pérez-Almendros, Luis Espinosa Anke, and Steven Schockaert. 2022. Semeval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. *arXiv preprint arXiv:2011.08320*.

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*, 55:477–523.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models.

Maximilian Schmidhuber. 2021. Universität regensburg maxs at germeval 2021 task 1: Synthetic data in toxic comment classification. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 62–68.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Nina Seemann, Yeong Su Lee, Julian Höllig, and Michaela Geierhos. 2023. Generalizability of abusive language detection models on homogeneous german datasets. *Datenbank-Spektrum*, 23(1):15–25.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 1–19. Springer.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Gabriel Louis Tan, Adrian Paule Ty, Schuyler Ng, Denzel Adrian Co, Jan Christian Blaise Cruz, and Charibeth Cheng. 2022. Using synthetic data for conversational response generation in low-resource settings. *arXiv preprint arXiv:2204.02653*.

Martin A Tanner and Wing Hung Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, and Surangika Ranathunga. 2018. Handling rare word problem using synthetic training data for sinhala and tamil neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Almira Osmanovic Thunström and Steinn Steingrimsson. 2022. Can gpt-3 write an academic paper on itself, with minimal human input?

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Minh-Thang Luong, Quoc V Le, Grady Simon, and Mohit Iyyer. 2021. Strata: Self-training with task augmentation for better few-shot learning. *arXiv preprint arXiv:2109.06270*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Dewayne Whitfield. 2021. Using gpt-2 to create synthetic data to improve the prediction performance of nlp machine learning classification models. *arXiv preprint arXiv:2104.10658*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340.

Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. 2016. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE.

Tomer Wullach, Amir Adler, and Einat Minkov. 2020. Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*, 25(2):48–57.

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.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to

design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

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). *arXiv preprint arXiv:1903.08983*.

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). *arXiv preprint arXiv:2006.07235*.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, 30(1):75–89.