


Don't Augment, Rewrite?

Assessing Abusive Language Detection with Synthetic Data

Camilla Casula , Elisa Leonardelli , Sara Tonelli 
ccasula@fbk.eu, eleonardelli@fbk.eu, satonelli@fbk.eu

 Fondazione Bruno Kessler, Italy
 University of Trento, Italy

Abstract

Research on abusive language detection and content moderation is crucial to combat online harm. However, current limitations set by regulatory bodies and social media platforms can make it difficult to share collected data. We address this challenge by exploring the possibility to replace existing datasets in English for abusive language detection with synthetic data obtained by rewriting original texts with an instruction-based generative model. We show that such data can be effectively used to train a classifier whose performance is in line, and sometimes better, than a classifier trained on original data. Training with synthetic data also seems to improve robustness in a cross-dataset setting. A manual inspection of the generated data confirms that rewriting makes it impossible to retrieve the original texts online.

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

1 Introduction

Abusive language detection¹ online has been a topic of interest to the NLP research community for the last ten years, with a number of datasets retrieved from social media platforms and annotated to be used for training and evaluation (Vidgen and Derczynski, 2020; Poletto et al., 2020). However, more and more restrictions are now being set that limit resharing of this kind of data even for research purposes (e.g. recent changes in X/Twitter terms of use). If data is used for commercial purposes, even more restrictions apply, while national and international organizations are enforcing regulations to protect users' privacy and limit data transfer, like the European General Data Protection Regulation.² Datasets created to train abusive language detection

¹In this work, we use *abusive* as an umbrella term to encompass both implicit and explicit attacks, including *hate speech* and *offensive language* (Caselli et al., 2020).

²<https://gdpr.eu/tag/gdpr/>

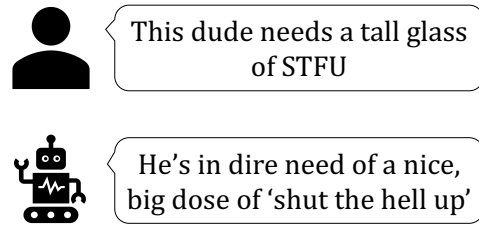


Figure 1: Original example and its corresponding synthetic rewriting.

systems deserve particular attention in this respect, since they are extremely useful to develop systems that contrast online hate, but on the other hand could be easily employed to profile users and target them. As discussed in Jahan and Oussalah (2023), even when such datasets do not contain user information, a search engine could be straightforwardly used to trace back the person who posted a certain message, nullifying anonymization efforts. Furthermore, in several jurisdictions around the world, social media users should be granted the so-called 'right to be forgotten', i.e. they can ask for their personal data, including data from social media posts, to be removed from platforms and, ideally, derived datasets.

In the light of these restrictions, it is likely that the availability of data to pursue research on abusive language detection and content moderation will represent an issue in the near future. We therefore address the following question in this work: *would it be possible to replace existing datasets for abusive language detection with synthetic ones by maintaining the same classification performance?* Performing this task with a good accuracy would present several advantages. For example, it would be possible to freely share datasets without the risk of disclosing user information or infringing terms of use and regulations. Furthermore, it would

mitigate the problem of data degradation, which makes social media datasets unusable after few years from release, since hateful content is frequently deleted (Klubička and Fernandez, 2018; Chung et al., 2019). Beside the benefits in terms of privacy, high-quality synthetic data could also reduce the need for human effort in dataset annotation, providing an effective way to automatically create large amounts of training data.

We present a first set of experiments in this direction by using a generative language model to rewrite two different abusive language datasets, comparing two prompt types. We then evaluate the quality of the generated data by using it for training abusive language detection classifiers.

2 Background

The use of synthetic data has been proposed as a way to mitigate some of the known issues with hate speech datasets (Vidgen and Derczynski, 2020), such as data scarcity (Founta et al., 2018) and negative psychological impact on annotators (Riedl et al., 2020). One of the applications of synthetic data is data augmentation (DA), in which new data is generated starting from a small set of gold instances, and which is mainly meant to address low-resource settings. Although DA based on large language models (LLMs) has been found to potentially lead to performance improvements for many NLP tasks (Feng et al., 2021; Chen et al., 2023), it has shown mixed results with regards to abusive language identification (Casula and Tonelli, 2023). Other work has focused on using fully synthetic data for training models in a few-shot and zero-shot setting (Li et al., 2023), but models trained on synthetic data were found to achieve worse performance than models trained on gold data, especially for subjective tasks. Differently from DA and the above work, we propose a *rewriting* approach, meant to address a scenario where enough gold data has been collected and annotated in the past, but cannot be reshared.

3 Data

In this work, we experiment with the creation of synthetic data starting from two different English datasets revolving around abusive language. First, the Multi-Domain Agreement dataset (MDA) by Leonardelli et al. (2021), which contains tweets annotated for offensive language with different inter-annotator agreement levels, spanning across

three topics: the Black Lives Matter movement, Covid-19 and the 2020 US elections. We use the set of this dataset containing 2,700 tweets, balanced across agreement levels, labels, and topics. Second, we use the Measuring Hate Speech (MHS) corpus (Kennedy et al., 2020; Sachdeva et al., 2022), which contains social media posts from multiple platforms annotated for hate speech. We use $\sim 10\%$ of this corpus for training (3,013 examples) given its large size, to reflect the size of the MDA dataset. We then test each model on both datasets to evaluate cross-dataset generalization. Furthermore, we also test our models on the **HateCheck** test suite (Röttger et al., 2021), a set of functional tests for finding specific weaknesses of hate speech detection models. For all datasets, we first carry out a pre-processing step in which we remove all user mentions and URLs, replacing them with ‘@USER’ and ‘URL’ respectively. Further details about the splits we use are reported in Appendix A.

4 Methods

Previous works experimenting with generative LLMs to augment abusive language datasets mostly exploit fine-tuning of generative models on existing gold data (Anaby-Tavor et al., 2020; Kumar et al., 2020), trainable components for task-specific decoding (Hartvigsen et al., 2022), or humans in the loop to evaluate generated sequences (Fantón et al., 2021; Chung et al., 2023). However, with the growing size of generative LLMs, making them more expensive to fine-tune, and their zero-shot capabilities thanks to instruction tuning, these models can often carry out numerous tasks without requiring any further fine-tuning (Wei et al., 2022). Because of this, we use a freely available instruction-tuned model in our experiments, Llama-2 chat 7B (Touvron et al., 2023), through the HuggingFace library (Wolf et al., 2020).³

The now widespread use of instruction-tuned generative large language models has also led to numerous efforts towards *alignment*, ideally in order to minimize inappropriate, offensive or unethical uses (Rao et al., 2023). While this is often preferable for many applications, it can make the creation of synthetic abusive language detection datasets complex, as these models are tuned to avoid producing abusive content due to its potentially harmful uses. Because of this, we frame the task not as

³<https://huggingface.co/meta-llama/Llama-2-7b-hf>

the *generation* of new, unseen data, but rather as the *rewriting* of existing gold sequences, so that *i*) the synthetic sequences are semantically close to existing data, inspired by the simple changes applied by Easy Data Augmentation (Wei and Zou, 2019), and *ii*) the synthetic sequences cannot in principle be traced back to any existing social media posts or their original posters.

4.1 Rewriting Original Examples

We frame our rewriting task as two established NLP tasks: **paraphrasing** and **formality style transfer**, with the aim of forcing rewriting to be informal, so that we counteract the tendency of aligned models to use language that is as neutral as possible.

In order to maximize our chances of producing at least one ‘valid’ sequence corresponding to each gold example, we produce multiple synthetic sequences. In particular, starting from each original text, we prompt Llama-2 with three different prompts for each task type, giving the model the original text and instructing it to reword it. For instance, for the *formality style transfer* task framing, one of our prompts would be:

*Rewrite this message more informally,
keeping the same meaning: “text”*
Rewritten text: ”

A full list of the prompts can be found in Appendix C. In addition to using 3 different prompts for each task framing, we run each prompt 3 times for each corresponding gold example, resulting in at least 9 synthetic texts for each source text for each prompt type.⁴ For generation, we use $top-p = 0.9$, a temperature of 1.0, and we set the minimum and maximum lengths of the generated sequences to 3 and 500, respectively.⁵

4.2 Filtering

Since we aim at obtaining synthetic data that *i*) cannot be reconnected to their source text but *ii*) preserves the original labels of the source data, we perform two filtering steps. First, we discard the synthetic sequences that are *verbatim* or extremely similar repetitions of the original gold texts using

⁴In some cases the model will continue producing paraphrases until it hits our maximum length, often resulting in more than 9 synthetic sequences being produced.

⁵The remaining hyperparameters we use are the default ones of the GenerationConfig HuggingFace class.

the TheFuzz library⁶, a Levenshtein distance-based tool to calculate string similarity. Then, we further filter the synthetic sequences using a classifier, following an established practice in generation-based data augmentation (Anaby-Tavor et al., 2020; Casula and Tonelli, 2023) to minimize issues with data preservation (Kumar et al., 2020). In particular, we train a Roberta Large classifier (Liu et al., 2019) on the original gold data, and then use this classifier to infer the class of the synthetic instances. We discard all sequences for which the predicted label of the synthetic text does not match the label of the original text used to create it. Finally, out of all the remaining sequences, we pick a random one to use as the synthetic equivalent of the original one in our experiments. If for a given original example no synthetic texts pass the filtering stage, we move onto the next example. As a result, the size of the original gold dataset tends to be bigger than the synthetic one. The total number of synthetic texts that pass filtering for each prompting type are reported in Table 1 in the $n(train)$ column.

Out of the synthetic texts that do not pass the filtering step, an overwhelming majority of them (between 95 and 98% across both datasets and prompting strategies) does not make it due to inconsistent label assignment, i.e. the classifier predicted a different label for the synthetic text than the label of the original. The remaining texts that are discarded during filtering are mostly almost-exact matches with the original data, and instances that do not pass either filter are extremely rare.

5 Evaluation

Since our main research question focuses on the possibility to replace abusive language datasets with synthetic data by maintaining the same performance level, we train classifiers on synthetic data derived from MDA and MHS and evaluate them on the gold test sets of both, including cross-dataset testing to assess model robustness. Additionally, we test our models with the HateCheck test suite. For our classification experiments, we fine-tune a Roberta Large classifier (355M parameters) (Liu et al., 2019) on the original and synthetic data for both gold datasets and for both prompting types. We select this model because it was the best performing one on the MHS dataset, as reported in Kennedy et al. (2020), and it outperformed BERT

⁶<https://github.com/seatgeek/thefuzz>. We discard all sequences scoring over 75 in terms of similarity with their original counterparts.

Test data →		MDA		MHS		HateCheck		
Training data ↓	$n(\text{train})$	M-F ₁	Ab-F ₁	M-F ₁	Ab-F ₁	M-F ₁	Ab-F ₁	
MDA	Original gold	2,161	0.779 \pm .009	0.720 \pm .008	0.661 \pm .011	0.595 \pm .007	0.519 \pm .021	0.573 \pm .033
	Synth: Paraphrase	1,444	0.779 \pm .013	0.706 \pm .022	0.680 \pm .005	0.607 \pm .004	0.508 \pm .009	0.552 \pm .016
	Synth: Formality	1,557	0.783 \pm .003	0.713 \pm .004	0.684 \pm .007	0.611 \pm .005	0.470 \pm .015	0.490 \pm .029
MHS	Original gold	3,013	0.540 \pm .034	0.260 \pm .063	0.791 \pm .006	0.688 \pm .010	0.338 \pm .029	0.206 \pm .049
	Synth: Paraphrase	2,435	0.629 \pm .014	0.423 \pm .025	0.787 \pm .005	0.694 \pm .008	0.351 \pm .025	0.236 \pm .051
	Synth: Formality	2,587	0.606 \pm .019	0.381 \pm .035	0.793 \pm .003	0.697 \pm .003	0.359 \pm .008	0.255 \pm .017

Table 1: Average results over 5 runs in terms of Macro-F₁ (M-F₁) and Abusive-class F₁ (Ab-F₁) \pm stdev. Grey cells denote out of distribution / cross-dataset performance. The $n(\text{train})$ column indicates the number of initial examples for gold data and the number of synthetic instances that passed filtering and are therefore used for training.

on MDA (Leonardelli et al., 2021). Furthermore, previous work showed that no relevant differences among BERT-like models could be observed with respect to the impact of synthetic data on model performance for hate speech detection (Casula and Tonelli, 2023). Further details are reported in Appendix B.

Classification Results Classification results of models trained on synthetic data are reported in Table 1. We report the mean macro-F₁ and abusive-class F₁ across 5 runs with different data shuffles and different model initializations, as well as the standard deviation across runs.

We observe that models trained on synthetic data *tend to perform similarly to the models trained on the original gold data, in some cases even with mild improvements*. This is in contrast with previous findings showing that synthetic data are generally not very helpful for subjective tasks such as this one (Li et al., 2023). However, this difference might be due to the fact that, for instance, Li et al. (2023) frame the creation of synthetic data as generation, not rewriting of existing examples, and they do not carry out any filtering on the artificial texts. Surprisingly, our models based on synthetic data perform well even if the training set size is smaller than the gold one. Furthermore, we observe improvements with regards to cross-dataset performance, especially in the case of the synthetic data produced starting from the MHS corpus. Indeed, the model trained on data rewritten from MHS yields an improvement over training using gold data when tested both on MDA and HateCheck, up to 16 abusive-class F₁ points on the former. These results suggest that synthetic data can potentially improve robustness in out-of-distribution scenarios, probably because lexical cues specific to the training data may be removed through rewriting,

		TTR	MTLD
MDA	Original gold	0.86	52.37
	Synth: Paraphrase	0.88	59.86
	Synth: Formality	0.88	69.70
MHS	Original gold	0.88	52.08
	Synth: Paraphrase	0.87	65.18
	Synth: Formality	0.87	66.46

Table 2: Lexical diversity measures on the original and synthetic data for both datasets.

achieving better generalization capabilities. Overall, neither prompting type appears to clearly outperform the other.

Lexical Diversity To further analyze differences between original and synthetic data, we compare their lexical diversity using Type Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005), which are calculated with the TAALED library⁷ on texts tokenized with Spacy.⁸ Results are reported in Table 2. While TTR is comparable on all datasets, generally indicating high lexical variability, MTLD shows a difference between gold and synthetic data. Indeed, synthetic data exhibits a higher degree of lexical diversity, especially if generated with the *formality* prompt type. The different output between TTR and MTLD may be due to the fact that the latter is more robust with regards to text length variations (Fergadiotis et al., 2015).

Lexical Analysis In addition to quantitatively measuring lexical diversity, we also inspect the data for any lexical cues that might influence generalization. Using the Variationist tool (Ramponi

⁷<https://pypi.org/project/taaled/>

⁸<https://spacy.io/>

et al., 2024), we calculate `npw_relevance`, a normalized class relevance metric based on PMI as seen in Ramponi and Tonelli (2022). Looking at what tokens tend to be more informative for each class, we consult the list of tokens whose informativeness for the abusive class changes the most in the gold and generated data. For the MDA dataset, the 5 most informative tokens for the abusive class are ‘CHEATER’, ‘Fuck’, ‘shit’, ‘ass’, and ‘fucking’. Conversely, in the synthetic data rewritten through paraphrasing, the most informative tokens for this class are ‘person’, ‘Fuck’, ‘individual’, ‘idiot’, and ‘expressing’, which seems to support the hypothesis that synthetic data might influence the reliance of models on lexical cues. We plan to further investigate this in future work.

Manual Inspection We finally perform a manual inspection of the generated data, to assess whether the synthetic data can be traced back to the original post through online search. We provide an annotator with 100 synthetic examples from the MDA dataset for each prompting type.⁹ Out of the original posts that are still online, 60 to 62% were found by the annotator through a search engine. However, *none of the original posts could be found starting from their synthetic counterparts*, showing the potential of this type of approach for data anonymization.

6 Conclusion

In this work, we perform a first exploration of abusive language detection using synthetic data generated through rewriting. We show that this is a promising research direction, since models trained on synthetic data can achieve a classification performance on par with models based on gold data, and even show better robustness in some cross-dataset settings. Furthermore, it was not possible to trace back the original data starting from the synthetic ones, even through a manual search online. Rewriting original texts seems to be an effective strategy both through paraphrasing and formality style transfer. We believe this approach to be a step forward for the development of datasets and systems for subjective tasks that are more privacy-aware and comply with existing regulations on personal data sharing, anonymization, and right to be forgotten.

⁹We only manually analyze MDA-derived data since we have the original Tweet IDs of the messages, to effectively check whether the original messages still exist online.

Limitations

Our method for creating synthetic data relies on the availability of pre-trained LLMs for generation, which could in some cases inject private information into the synthetic sequences. In this respect, it would be interesting to investigate in the future a possible integration between our approach and differentially private models (Yu et al., 2022; Matzken et al., 2023) to mitigate this kind of risk.

All our paraphrasing experiments rely on Llama-2 chat 7B, which we select because it is widely used in research and it is freely available. We use a single model to carry out both types of rewriting to avoid confounding effects on our results. However, using other LLMs to paraphrase would allow us to test more thoroughly the robustness of our approach. We plan to experiment with more models in future work.

Our experiments present a first exploration towards the use of synthetic data for abusive language detection using a privacy-aware approach. Our approach may be comparable or even better than training a system with real data in terms of performance. However, we acknowledge that the insights, domain and linguistic knowledge coming from real data represent a crucial contribution to better understand online communication, and this would not be possible with synthetic data.

Ethical Considerations

The main motivation behind our work is the need to pursue research on abusive language detection that complies with privacy-preserving principles and regulations. However, our approach does not guarantee that personal information contained in a source text are erased during rewriting. Personal information leaked from the generative model could be even introduced during rewriting. Although we have not found such cases during the manual inspection of our synthetic data, we advocate for a combination between our approach with differential privacy. In addition, some biases could be propagated through rewriting, so further research in this direction would be needed to assess this type of risk.

Finally, all the ‘original’ examples used in this paper were slightly modified by hand, so that they cannot be traced back to any existing posts.

Acknowledgements

This work has been funded within the European Union’s ISF program under grant agreement No. 101100539 (PRECRISIS). We also acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU.

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.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. [I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language.](#) In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.
- 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.
- 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.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. [CONAN - COUNTER NARRATIVES THROUGH NICHE SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, 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 Edward 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.
- Gerasimos Fergadiotis, Heather Harris Wright, and Samuel B Green. 2015. Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58(3):840–852.
- 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.](#) *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).
- 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.
- Md Saroar Jahan and Mourad Oussalah. 2023. [A systematic review of hate speech automatic detection using natural language processing.](#) *Neurocomputing*, 546:126232.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. [Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application.](#) ArXiv:2009.10277 [cs].
- Filip Klubička and Raquel Fernandez. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- 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*.
- Cleo Matzken, Steffen Eger, and Ivan Habernal. 2023. [Trade-offs between fairness and privacy in language modeling](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6948–6969, Toronto, Canada. Association for Computational Linguistics.
- Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. [Resources and benchmark corpora for hate speech detection: a systematic review](#). *Language Resources and Evaluation*, 55:477 – 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: System Demonstrations (To Appear)*. Association for Computational Linguistics.
- 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.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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](#).
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pier-

ric 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.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarini, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*.

A Data Details

We use the default train-dev-test splits of the MDA dataset, with 2,160 class-balanced training examples, 540 development examples, and 3,057 test examples. For the MHS dataset, we use the *hate speech* label rather than the hate speech scores, since we focus on binary classification. The hatespeech label in the MHS dataset can assume three values (0: *non hateful*, 1: *unclear*, 2: *hateful*), we average all the annotations for a given post, mapping it to *abusive* if the average score is higher than 1 and to *non abusive* if it is lower. After this process, we are left with 35,243 annotated posts, of which 9,046 are annotated as containing hate speech ($\sim 26\%$). We use a subset of this dataset in order to have a comparable size with the other dataset we use, selecting 10% of the dataset as the test set (3,524), 5% of the remaining examples as development data (1,586 examples) and finally 10% of the remaining texts as the training data (3,013 posts). The random seed we use for splitting¹⁰ and shuffling the data for all datasets is 2023.

B Implementation Details

We use the HuggingFace library for all model implementations. For the RoBERTa classifiers, we use the default hyperparameters of the TrainingArgs class, setting batch size to 16, the maximum sequence length to 150, and the learning rate to 5e-6. For the models trained on the original data, we run training for 10 epochs, selecting the epoch with the lowest validation loss. After a manual analysis of the best epochs in most of the runs with the original data, we pick 3 epochs for training the models on

¹⁰With the `train_test_split` function from `sklearn.model_selection`.

synthetic data, as using synthetic development data would be misleading, and we aimed at not using any gold data during the training phase. Generation took about 180 hours and training of classifiers took about 10 hours on a Nvidia A40 GPU.

C Llama-2 Chat Prompts

In this section we provide the full prompts we use for Llama-2 chat. For the **paraphrasing** task framing, we use the following prompting templates:

1. *Paraphrase this text:* “text”
Paraphrased text: “
2. *Reword this text, preserving meaning and tone:* “text”
Reworded text: “
3. *Rewrite this message keeping the same meaning:* “text”
Reworded text: “

For the **formality style transfer** task framing we use the following templates for prompting:

1. *Paraphrase this text in a more informal way:* “text”
Paraphrased text: “
2. *Reword this text, preserving meaning and tone but using more informal language:* “text”
Reworded text: “
3. *Rewrite this message more informally, keeping the same meaning:* “text”
Reworded text: “