Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset

Janis Goldzycher[1] **Paul Röttger**[2] **Gerold Schneider**[1]
[1]University of Zurich, Zurich, Switzerland
[2]Bocconi University, Milan, Italy

Abstract

Hate speech detection models are only as good as the data they are trained on. Datasets sourced from social media suffer from systematic gaps and biases, leading to unreliable models with simplistic decision boundaries. Adversarial datasets, collected by exploiting model weaknesses, promise to fix this problem. However, adversarial data collection can be slow and costly, and individual annotators have limited creativity. In this paper, we introduce GAHD, a new German Adversarial Hate speech Dataset comprising ca. 11k examples. During data collection, we explore new strategies for supporting annotators, to create more diverse adversarial examples more efficiently and provide a manual analysis of annotator disagreements for each strategy. Our experiments show that the resulting dataset is challenging even for state-of-the-art hate speech detection models, and that training on GAHD clearly improves model robustness. Further, we find that mixing multiple support strategies is most advantageous. We make GAHD publicly available at https://github.com/jagol/gahd.

Content Warning: This paper contains illustrative examples of hate speech.

1 Introduction

Robust hate speech detection is essential for addressing and analyzing online hate on a large scale. Hate speech detection models are typically trained on datasets sourced from social media or newspaper comment sections (Poletto et al., 2021). However, such datasets are known to have systematic gaps and biases, which leads to models that suffer from lexical overfitting and poor generalisability (Vidgen et al., 2019; Wiegand et al., 2019; Poletto et al., 2021; Röttger et al., 2021).

Dynamic adversarial data collection (DADC), seeks to address this issue, by tasking annotators to create texts that trick a model, the *target* model, into incorrect classifications (Kiela et al., 2021).

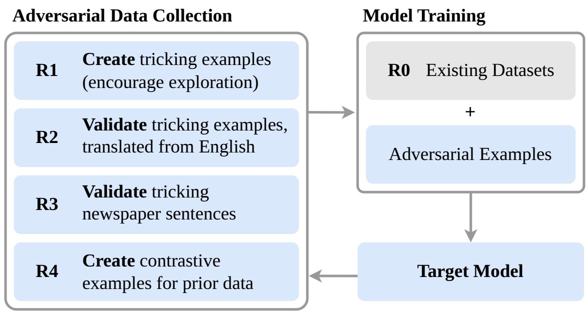

Figure 1: We use four rounds of **dynamic adversarial data collection** (Kiela et al., 2021) to improve a German hate speech classifier. We start with a target model trained on existing datasets. Then, in each round (R1-R4), annotators try to trick the target model using a different method. After each round, we train a new target model including the new adversarial examples.

The newly-created data is added to the training data, and the target model is then retrained on all data, making it more robust. This process is repeated across multiple rounds. Vidgen et al. (2021) use DADC to create an English hate speech dataset, and show that training on their data substantially improves model robustness. However, DADC is time-consuming, expensive, and can result in a homogenous dataset, unless annotators explore diverse strategies for tricking the target model. In this paper, we introduce GAHD, a new **G**erman **A**dversarial **H**ate speech **D**ataset, collected with four rounds of DADC. To address the limitations of prior DADC work, we use a new strategy in each round to support annotators in finding diverse adversarial examples, in a time-efficient manner. Figure 1 shows our improved DADC process: In R1, the first round, we let annotators come up freely with their own adversarial examples. For R2, we provide the annotators with English-to-German translated adversarial examples as candidates to validate or reject, and as a way to inspire new, derived examples. In R3, annotators validate sentences from German newspapers that the target model

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4405–4424

labeled as hate speech. Due to their origin, it is unlikely that these sentences are hate speech, which makes them likely adversarial examples. For R4, we task annotators with creating contrastive examples by modifying previously collected examples in a way that flips their label.

GAHD contains 10,996 adversarial examples, with 42.4% labeled as hate speech. 1,300 entries are paired with a contrastive example. Evaluating the target model after each round demonstrates large improvements in model robustness, with almost 20 percentage point increases in macro F_1 on the GAHD test split (in-domain), and German HateCheck test suite (out-of-domain) (Röttger et al., 2022). We further evaluate the contribution of individual rounds, while controlling for data size, observing that rounds with manually-crafted examples are more effective, but that mixing multiple rounds with different data collection strategies leads to more consistent improvements. Finally, we benchmark a range of commercial APIs and large language models (LLMs) on GAHD, finding that the APIs generally struggle, with only GPT-4 achieving over 80% macro F_1. In summary, our contributions are:

1. We introduce GAHD, the first German Adversarial Hate speech Dataset, containing ca. 11k examples collected by DADC.
2. We propose new strategies for collecting more diverse adversarial examples in a more time-efficient manner, thus improving DADC.
3. We demonstrate the usefulness of GAHD for improving model robustness, and evaluate the contribution of individual rounds.
4. We benchmark a range of commercial APIs and LLMs on GAHD.

2 Annotation

2.1 Annotation Setup

We collect adversarial examples with binary annotations – *hate speech* or *not hate speech* – using the Dynabench platform (Kiela et al., 2021). Dynabench provides an interface for dynamic adversarial data collection. Annotators enter self-created examples via the interface along with what they consider to be the correct label. The target model then predicts a label and the annotator is shown if the predicted label agrees with the provided label or disagrees with it. All entered examples are validated once by another annotator and, in case of disagreement, forwarded to an expert annotator, who makes a final decision. The paper authors take the role of expert annotator.

2.2 Definition of Hate Speech

There is no universally accepted definition of hate speech. For this paper, we follow the majority of recent work and define hate speech as follows: For an utterance to be categorized as hate speech, abusive or discriminatory language must be directed either at a protected group or at an individual specifically as a member of a protected group (Poletto et al., 2021; Yin and Zubiaga, 2021). The term "protected groups" can be interpreted as referring either to all social groups defined via characteristics such as race, religion, gender, sexual orientation, disability, and similar or only marginalized groups defined via these characteristics (Khurana et al., 2022). For this work, we only consider marginalized social groups as protected groups. Further, we deviate from previous definitions, by including *poor people* as a protected group, as has been argued for by Kiritchenko et al. (2023).

2.3 Annotation Guidelines

We follow a prescriptive approach to annotation (Rottger et al., 2022), giving annotators detailed instructions and training to apply our annotation guidelines. Before R1, the annotators received in-person annotation instructions including a presentation and discussion session on what is considered hate speech in this dataset. In addition to a detailed definition of hate speech the instructions contain three main points: (1) They specifically emphasize that hate speech depends on cultural context, making annotators aware of how protected groups and stereotypes in a German context might differ from protected groups, in a different cultural context. (2) The goal of GAHD is to cover protected groups, controversial issues, and stereotypes of all three major German-speaking countries (Austria, Germany, and Switzerland). (3) Annotators should aim for examples that clearly fall into either hate speech or not-hate speech, and avoid exploiting the definitional grey area.

2.4 Annotator Details

To support diverse model-tricking strategies, we distributed the annotation load between as many annotators as was possible given budget limitations and administrative contraints. We recruited seven annotators for 30 hours of work each. All annotators are students or work at a university. All annotators are native or highly competent German speakers with basic to advanced knowledge of computational linguistics. Three of the annotators had prior specific knowledge about hate speech detection gained through courses or student projects. For R4, we used the remaining funds to hire two additional annotators. We compensated all annotators well above the minimum wage, according to university guidelines, taking into account their academic degrees. Appendix F contains a data statement (Bender and Friedman, 2018) with additional details.

3 Dynamic Adversarial Data Collection

3.1 Target Model

As our target model across all rounds, we use gelectra-large, a German Electra large model with ca. 335m parameters, which outperforms other similarly-sized German and multilingual models on German text (Chan et al., 2020).[1] We chose this model because it is both strong and lightweight, so that annotators receive fast feedback (model tricked / not) on the examples they create.

To train an initial target model for R1, we fine-tuned gelectra-large on training splits of five German hate speech detection datasets with similar hate speech definitions or related labels that can be mapped to our definition of hate speech: DeTox (Demus et al., 2022), the German part of HASOC 2019 SubTask 2 (Mandl et al., 2019), the German part of HASOC 2020 Subtask 2 (Mandl et al., 2021), and the RP-Crowd dataset (Assenmacher et al., 2021). We divided all datasets randomly into training (70%), development (15%), and test (15%) splits. After each round of DADC, we split the newly collected data using the same ratios and added it to the existing splits. Further details about the initial datasets and model training are available in Appendices B and C.

3.2 Round 1: Unguided Data Creation

For R1, we tasked annotators to fool the target model in the Dynabench interface without further guidance. Annotators entered 2,209 examples, with 45.3% being hate speech. We found 34 duplicates leading to 2,175 unique examples. Each example was validated once, leading to a Cohen's Kappa of 0.83. There were 208 disagreements, which we resolved via expert annotation by one of the paper authors.

[1] huggingface.co/deepset/gelectra-large

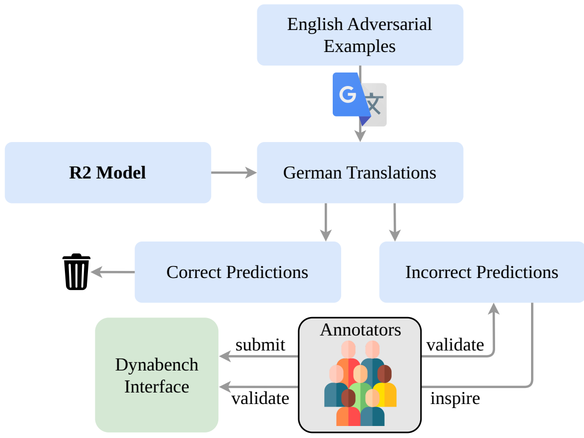

Figure 2: DADC workflow for R2, where we let annotators validate model tricking translations of English adversarial examples.

Lessons We observe that many disagreements in R1 stem from three main issues: 1) Definition of protected migrant groups: Initially, there was confusion about whether all migrants, including those from Western countries such as the U.S. and France, should be considered protected groups by virtue of being migrants. We specified the annotation guidelines such that only migrant groups with a history of marginalization or discrimination in German-speaking countries are classified as protected. 2) Author's stance towards quoted speech: Some examples included quotes of or references to hate speech without any indication of the author's view on it. Since the author's position (supporting or against the referenced hate speech) is essential in determining if a text is hate speech, and with the motivation of avoiding noise, we now ask annotators to include subtle hints of the author's stance in their texts. 3) Ambiguity in targeting protected groups: There were instances where calls for violence or similar actions were made against unspecified social groups. Our revised guidelines specify that if the language indicates that any marginalized group (without needing to specify a specific protected group) is being targeted by vague calls for violence, the text should be classified as hate speech. Conversely, if there is no indication of targeting any protected group, it does not meet our hate speech criteria. To ensure that the already-validated R1 examples were in line with the refined guidelines, an expert annotator annotated the targeted groups in all R1 examples, and systematically adjusted labels per target group.

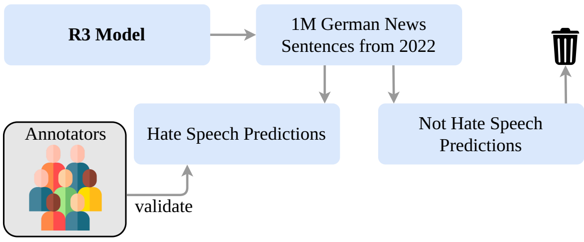

Figure 3: Workflow of R3, where we task annotators with validating model tricking newspaper sentences.

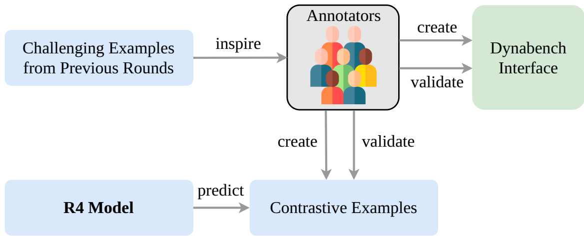

Figure 4: Workflow of R4, where we let annotators create contrastive examples to challenging entries from previous rounds.

3.3 Round 2: Translated Adversarial Examples

For R2, we translated English adversarial examples collected by Vidgen et al. (2021) to German using Google Translate[2] and let the target model – now additionally trained on R1 data – classify the examples. Examples where the model prediction disagreed with the original English dataset label became candidates for adversarial examples. Since it is possible that translating the examples introduced errors, or that the examples simply do not apply to the German-speaking context, we gave each example to an annotator for validation. Further, we gave annotators the option to enter examples that were inspired by examples encountered during validation in the Dynabench interface.

Overall, this led to 3,996 validated examples translated from English, with 74.4% labeled as hate speech. Further, the annotators entered and validated 138 new examples (43.5% hate speech) via the Dynabench interface, with a high Cohen's Kappa of 0.99. We attribute this high inter-annotator agreement to the high degree of submitted examples that are clearly hate speech or not.

Lessons During a manual inspection, we found instances where annotators accepted examples containing derogatory expressions, such as slurs that Google Translate did not translate from English to German. We adopt the annotator's reasoning that certain English slurs, like "n***a", or "c**t" have been integrated into German-speaking culture as Anglicisms. Therefore, we deem these untranslated slurs to be useful and keep them in GAHD.

3.4 Round 3: Newspaper Sentences

For R3, we used the sentences sampled from German newspaper articles published in 2022(Goldhahn et al., 2012).[3]Assuming that officially published news is unlikely to contain hate speech, any sentence classified as hate speech is likely a false positive and thus an adversarial example. We used the target model to classify one million news sentences, which yielded 8,056 classified as hate speech. We then sorted the flagged sentences by how confident the model was in its prediction and distributed them to annotators, with higher-confidence sentences being reviewed first. Overall, this resulted in 3,227 validated examples, with 87 annotated as hate speech. We removed three examples for containing metadata tags due to parsing errors. An expert annotator validated the only annotations marked as hate speech, disagreeing on 40 of the 87 examples. Inspecting the disagreements shows that they come from one annotator and mainly stem from two reasons: (1) labeling hate against non-protected groups as hate speech and (2) marking referenced but not endorsed hate speech as hate speech.

3.5 Round 4: Contrastive Examples

In R4, we focused on gathering contrastive examples for particularly challenging entries from previous rounds. We let the target model predict on data gathered in the previous rounds and collected all incorrect predictions as well as correct predictions that were made with high uncertainty. We then gave each of the nine annotators ca. 300 of these examples, and tasked them with modifying the given example to flip the label from hate speech to not-hate speech and vice versa. Instead of providing a modified, contrastive example, annotators also had the option to *disagree* with the label of the given example, *flag* the given example, or *skip* if the example is unsuitable for a contrastive example. Overall, we collected 1,253 contrastive examples (36.8% hate speech), and 132 *disagree*, and 154

[2] https://translate.google.com

[3] The data can be downloaded here: https://wortschatz.uni-leipzig.de/de/download/German#deu_news_2022

Round	Hate	No Hate	Total
R1	1,000 (46.0%)	1,175 (54.9%)	2,175
R2	3,043 (73.6%)	1,091 (26.4%)	4,134
R3	48 (01.5%)	3,179 (98.5%)	3,227
R4	575 (39.4%)	885 (60.6%)	1,460
Total	4,666 (42.4%)	6,330 (57.6%)	10,996

Table 1: Number of examples in GAHD across rounds.

Split	Hate	No Hate	Total
Train	3,265 (42.4%)	4,436 (57.6%)	7,701
Dev	709 (43.0%)	940 (57.0%)	1,649
Test	692 (42.0%)	954 (58.0%)	1,646
Total	4,666 (42.4%)	6,330 (57.6%)	10,996

Table 2: Label distribution in GAHD across data splits.

flag annotations. An expert annotator validated all contrastive examples, leading to a Cohen's Kappa of 0.89. The expert annotator also resolved the *disagree* and *flag* annotations.

Annotators primarily flagged examples for being incomplete, or very vague sentences so that a clear meaning is hard to assign. Almost all of those sentences were labeled as not-hate speech. Considering that a sentence without a clear meaning does not constitute hate speech, it can be a valid instance of not-hate speech. Therefore, we chose to keep these examples in our dataset and showcase a selection in Appendix D Table 7.

Annotators additionally entered and validated 160 new examples via the Dynabench interface, with a Cohen's Kappa of 0.89. On inspecting the R4 data from Dynabench, we observed that many examples were label-inverting perturbations of each other, effectively making them contrastive examples too.

3.6 Full Dataset

The final dataset contains 10,996 examples, with 4,666 (42.4%) labeled as hate speech. Table 1 shows a breakdown by round. After each round, we randomly split the collected data into training (70%), development (15%), and test split (15%) resulting in the distribution shown in Table 2.

Model Error Rate In R1, annotators successfully tricked the target model with 41.3% of examples. In R2, 34.5% of examples submitted via the Dynabench interface tricked the model. In R4, 37.8% of contrastive examples, and 31.3% of examples submitted via Dynabench tricked the model. Translated adversarial examples (R2) and newspaper sentences (R3) have a near 100% model tricking rate, since we only included them for having fooled the target model.

Inter-Annotator Agreement The inter-annotator agreement varied across rounds but was generally high. We speculate that the variation in agreement could stem from the fact that, not every annotator contributed equally in each round. If annotators, whose view on hate speech is more aligned, contributed more examples and validations in the same round, we achieve a higher agreement. Based on manual inspection we believe that in later rounds annotators produced examples that align more clearly with our definitions of either hate speech or not hate speech, making it less likely that annotators disagree on a label.

Clustering-Based Analysis To give a thematic overview, we cluster and visualize GAHD. Concretely, we embed all examples using `all-mpnet-base-v2` from the sentence transformers library (Reimers and Gurevych, 2019, 2020), reduce embedding dimensionality with UMAP (McInnes et al., 2020), and cluster the embeddings using HDBScan (Ester et al., 1996). Finally, we use GPT3.5-turbo[4] to generate cluster descriptions based on the top words (ranked via TF-IDF) and sentences of the cluster. We remove generic opening phrases from cluster descriptions, like "Cluster of texts [...]" or "Texts discussing [...]".

We obtain 22 clusters, ranging in size from ca. 60 examples to over 1,500. 3,700 examples remain uncategorized. Figure 5 shows the clusters projected onto two dimensions along with their cluster descriptions. Additionally, we provide an example for each cluster in Appendix D Table 8. We observe that the clustering leads to a categorization into protected groups and discourse topics such as COVID-19 (topic 1), the Russia-Ukraine war (topic 3) or football (topic 13). Further, the descriptions often highlight aspects about a protected group, indicating how texts target them. For example, the descriptions of the clusters 8, 9, and 11 suggest that these clusters revolve around immigrants having a perceived negative impact on social services and being a threat to national identity.

[4] https://platform.openai.com/docs/models/gpt-3-5

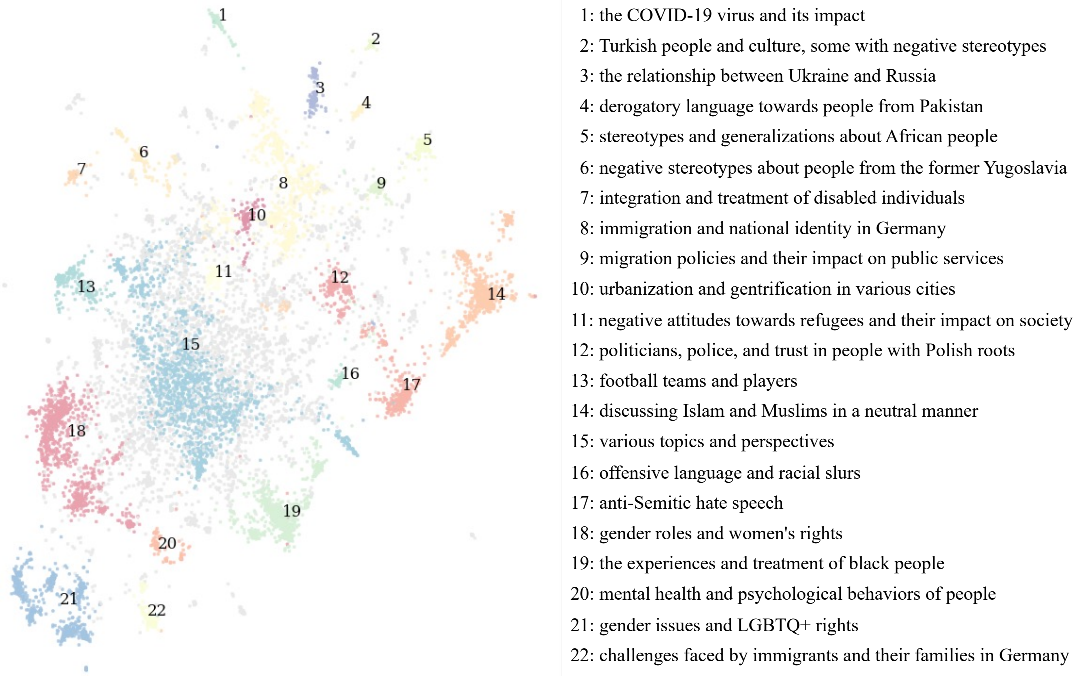

Figure 5: An overview of the most important topics in GAHD. We generate the topics via clustering and use GPT-3.5 to obtain cluster descriptions. Section 3.6 describes the procedure.

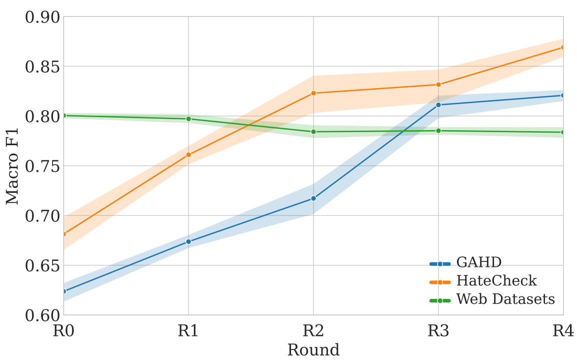

Figure 6: Target model performance on different test sets as we add new training data across four rounds of DADC.

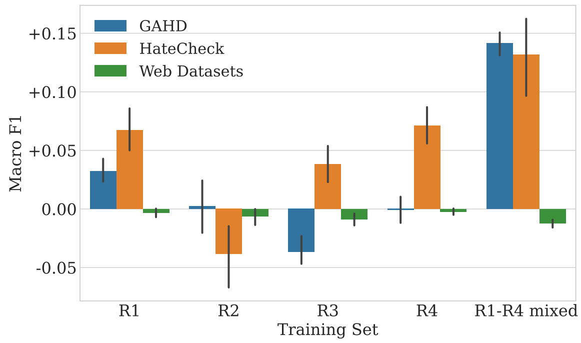

Figure 7: Impact on macro F_1 on different test sets when including 800 examples from a given round in the training data.

4 Experiments

4.1 Does GAHD Improve Model Robustness?

We want to test to what degree GAHD improves robustness systematically. For that purpose, we train gelectra-large on the web-sourced datasets from Section 3.1, and add the training splits of each round incrementally. We use macro F_1 to measure performance.

Evaluation Datasets We evaluate on the test split of GAHD, and on the combined test splits of the initial, web-sourced datasets described in Section 3.1. We further evaluate on the German part of HateCheck (Röttger et al., 2021, 2022), a synthetic test suite for model evaluation, and identification of critical model weaknesses.

Results Figure 6 displays the results averaged over ten random seeds. The shaded areas show the bootstrapped 95% confidence intervals around the average performance. Each new round clearly improves the performance on GAHD and HateCheck with earlier rounds having a larger impact than later rounds. On the web-sourced datasets the performance drops slightly, after including R2 data. Finally, including all GAHD rounds in the training ("R4") leads to an increase of 18 to 20 percentage points on GAHD and HateCheck.

Functionality	Label	R0	R1 800	R2 800	R3 800	R4 800	R1-R4 800	R1-R4 All
Expression of strong negative emotions (explicit)	H	0.993	-0.050	+0.007	-0.071	-0.164	-0.036	-0.050
Description using very negative attributes (explicit)	H	0.993	+0.000	+0.007	-0.021	-0.036	-0.057	-0.014
Dehumanisation (explicit)	H	0.979	+0.014	+0.021	-0.014	-0.043	+0.021	+0.021
Implicit derogation	H	0.745	-0.034	+0.234	-0.310	-0.303	+0.034	+0.159
Direct threat	H	0.936	-0.079	+0.057	-0.271	-0.379	-0.071	+0.000
Threat as normative statement	H	0.986	-0.143	+0.014	-0.157	-0.150	-0.121	-0.036
Hate expressed using slur	H	0.925	+0.017	+0.025	-0.133	+0.017	+0.058	+0.033
Hate expressed using profanity	H	0.971	-0.064	+0.029	-0.057	-0.071	-0.021	+0.014
Non-hateful use of profanity	NH	0.960	+0.040	-0.010	+0.040	+0.030	+0.040	+0.040
Hate expressed through reference in subsequent clauses	H	0.993	-0.064	+0.007	-0.214	-0.143	-0.029	+0.007
Hate expressed through reference in subsequent sentences	H	0.979	-0.043	+0.021	-0.143	-0.079	+0.007	+0.007
Hate expressed using negated positive statement	H	0.829	-0.057	+0.171	-0.193	-0.350	+0.029	+0.114
Non-hate expressed using negated hateful statement	NH	0.243	+0.307	-0.229	+0.386	+0.550	+0.550	+0.336
Hate phrased as a question	H	0.979	-0.129	-0.021	-0.207	-0.207	-0.264	-0.071
Hate phrased as an opinion	H	0.993	-0.064	-0.014	-0.093	-0.086	-0.093	-0.021
Neutral statements using protected group identifiers	NH	0.357	+0.464	-0.314	+0.450	+0.579	+0.550	+0.586
Positive statements using protected group identifiers	NH	0.243	+0.290	-0.233	+0.343	+0.571	+0.552	+0.529
Denouncements of hate that quote it	NH	0.290	+0.135	+0.032	+0.258	+0.368	+0.529	+0.413
Denouncements of hate that make direct reference to it	NH	0.374	+0.374	-0.039	+0.355	+0.400	+0.439	+0.284
Abuse at objects	NH	0.923	-0.062	-0.062	+0.015	+0.031	+0.031	+0.031
Abuse at individuals (not as member of a prot. group)	NH	0.723	+0.108	+0.031	+0.123	+0.062	+0.231	+0.246
Abuse at nonprotected groups (e.g. professions)	NH	0.615	+0.092	-0.062	-0.015	+0.015	+0.262	+0.338
Swaps of adjacent characters	H	0.964	-0.093	+0.036	-0.193	-0.164	-0.129	-0.143
Missing characters	H	0.907	-0.029	+0.071	-0.086	-0.057	-0.014	+0.036
Missing word boundaries	H	0.884	-0.013	+0.071	-0.090	-0.032	+0.006	+0.039
Added spaces between chars	H	0.477	-0.155	+0.400	-0.200	-0.187	-0.026	+0.226
Leet speak spellings	H	0.897	-0.123	+0.071	-0.155	-0.090	-0.013	+0.032
Full HateCheck		0.768	+0.028	+0.012	-0.021	+0.013	+0.098	+0.122

Table 3: Impact of including GAHD in the training data on the performance on individual HateCheck functionalities. The label "H" refers to hate speech and "NH" to non-hate speech. We mark accuracies below 0.7 on R0 in red.

Error Analysis We analyze how training on GAHD affects the performance on individual HateCheck functionalities, to gain insights into strengths and weaknesses introduced by GAHD. Table 3 column "R1-R4 All" shows the differences in performance after training on the full GAHD training set compared to only training on the web-sourced datasets ("R0"). Note, that we use accuracy scores since each functionality only contains one class, making macro F_1 unsuitable. We observe that the R0 model struggles on non-hate speech functionlities, such as processing of counter speech, non-hateful speech about protected groups, and abuse that is not targeted at protected groups. Including GAHD in the training data fixes these weaknesses.

4.2 Which Round Provided the Most Effective Examples?

To isolate the effect of each round and control for dataset size, we randomly sample 800 examples from the training split of each round and compare the effect of adding these to the training splits of the web-sourced data. In an additional scenario we draw 800 examples from the full GAHD training split, mixing all rounds. We use the same gelectra-large model and hyperparameters as in the previous section, and perform the experiments over ten random seeds for sampling as well as model training.

Results Figure 7 shows the results. We observe that the manually created examples from R1 and R4 have more positive effects on performance than the collected and validated examples from R2 and R3. Examples from these two rounds have mixed effects when used in isolation from the other rounds. However, combining data from all four rounds yields by far the best results, and clearly outperforms standard DADC as done in R1. This shows that introducing and combining support methods for annotators not only makes data creation more efficient, but can also increase the effectiveness of examples.

Error Analysis In Table 3, columns labeled "R1 800" through "R4 800" and "R1-R4 800" demonstrate the impact of including 800 examples from a specific round or from all rounds in the training data. We observe that R1, R3, and R4 have positive effects on the same functionalities, all containing non-hate speech. R2 impacts these functionalities negatively but has positive impacts on functionalities containing hate speech. We believe that the high amount of hate speech in R2 compared to the other rounds causes this behaviour.

Model	0-Shot	5-Shot
LeoLM 7B Chat	0.305	0.463
LeoLM 13B Chat	0.341	0.655
LeoLM 70B Chat	0.591	0.762
GPT-3.5	0.790	0.783
GPT-4	**0.809**	**0.833**
Content Moderation APIs		
Perspective		0.610
OpenAI		**0.695**
Target Model		
gelectra-large R0		0.623
gelectra-large R4		**0.822**

Table 4: Macro F_1 of LLMs and content moderation APIs on the GAHD test set. We include the results of gelectra-large, our target model, for comparison.

4.3 How Robust are Large Language Models and Commercial APIs?

To assess the difficulty of GAHD and to provide additional baseline results, we benchmark a range of LLMs and content moderation APIs on GAHD.

LLMs We evaluate the proprietary GPT-3.5 and GPT-4 language models.[5] (OpenAI, 2023) We also test the openly-available LeoLM models, which are based on Llama 2 (Touvron et al., 2023), and have been further pretrained and instruction tuned for German.[6] We evaluate all models in a zero-shot and five-shot scenario.

Content Moderation APIs The Perspective API by Google Jigsaw[7] and the content moderation API by OpenAI[8] both provide predictions, given an input text, for a range of attributes such as toxicity, or profanity. We use Perspective's predictions for the attribute *identity_attack*, and OpenAI's predictions for the attribute *hate*. Both attributes are defined via protected groups and closely align with our definition of hate speech. Appendix E contains additional evaluation details about LLM prompting and API usage.

[5]See: https://platform.openai.com/docs/models

[6]The creators of the LeoLM model suite have not yet released a paper. The training procedure is described in this blog post: https://laion.ai/blog/leo-lm/.

[7]https://www.perspectiveapi.com/

[8]https://platform.openai.com/docs/guides/moderation

Results As for the previous experiments, we evaluate with macro F_1 on the test split of GAHD. Table 4 shows the results. The GPT models achieve the highest scores, with GPT-4 being the only model that scores above 80%. LeoLM 7B obtains the lowest scores. Larger LeoLM Models achieve higher performances without reaching the GPT models. All LLMs except for GPT-3.5 benefit from examples in the prompt. The OpenAI API clearly beats Perspective API but falls behind the GPT models. Comparing these results to our fine-tuned gelectra models, we observe that fine-tuning on the train split of GAHD leads to the second highest scores, behind GPT-4 five-shot.

Error Analysis We focus on analyzing persistent errors where either both APIs or all LLMs in the zero-shot and five-shot scenarios predicted wrong. Persistent API errors make up 42% (315 examples) of all API errors. 67% of these errors belong to R2 and are mostly false negatives. In a manual analysis, we find that many of these false negatives contain group references that are hard to resolve such as camel-derived words to reference Arabic people or terms with modified spelling such as “chhhhinese”. There are 30 examples that all LLMs misclassified in both the 0-shot, and the 5-shot scenario. These are exclusively false positives. Many are counter-speech or reporting about hate crimes.

5 Related Work

Dynamic Adversarial Data Collection There is a growing body of work demonstrating that DADC improves the robustness and generalisability of NLP models on a wide range of tasks (Yang et al., 2017; Minervini and Riedel, 2018; Zellers et al., 2018; Dinan et al., 2019; Dua et al., 2019; Bartolo et al., 2020; Nie et al., 2020; Kiela et al., 2021). DADC further leads to datasets that are more syntactically and lexically diverse than non-adversarial data (Wallace et al., 2022). A branch of research building on this paradigm, exploring how DADC can be made more efficient, has shown that data augmentation for adversarial data improves model generalisation (Bartolo et al., 2021) and that supporting annotators by generating suggestions can improve the annotator efficiency and model tricking rate (Bartolo et al., 2022). Two previous papers applied DADC to hate speech. The first created an English hate speech dataset over four rounds of DADC (Vidgen et al., 2021). In contrast to our work, the authors relied on manually crafted examples and rule-based perturbations. The second paper uses DADC to create an English test suite for emoji-based hate speech (Kirk et al., 2022).

Hate Speech Datasets Hate speech detection datasets are typically sourced from social media, and are annotated on a post-level for binary or ternary classification (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020; Poletto et al., 2021). Sometimes more fine-grained annotations schemes are employed (Founta et al., 2018; Vidgen et al., 2019; Vidgen and Derczynski, 2020; Mollas et al., 2022). Adversarial datasets for hate speech can be categorized into collected web-sourced datasets (Sarkar and KhudaBukhsh, 2021), manually created datasets (Vidgen et al., 2021), and generated datasets (Cao and Lee, 2020; Hartvigsen et al., 2022; Ocampo et al., 2023). A range of adversarial attacks and perturbations on hate speech detection models have been proposed and analyzed (Gröndahl et al., 2018; Oak, 2019; Alsmadi et al., 2021; Grolman et al., 2022; Samory et al., 2021; Kumbam et al., 2023), leading to research on how to defend against such attacks (Moh et al., 2020). Finally, the goal of preventing models from relying on spurious correlations has motivated contrastive data augmentation (Gardner et al., 2020; Kaushik et al., 2020) and automatic counterfactual data augmentation for sexism and hate speech detection (Sen et al., 2022, 2023).

6 Conclusion

In this paper, we presented GAHD, a German Adversarial Hate speech Dataset produced via dynamic adversarial data collection (DADC). Across four rounds of data collection, we explored new strategies for supporting the annotators in efficiently creating diverse examples by suggesting candidates for validation or inspiration. In total, GAHD comprises 10,996 examples (42.4% hate speech), including 1,300 contrastive examples. Our experiments showed that (1) training on GAHD clearly improves the robustness of hate speech detection models, demonstrated by increases of 18-20 percentage points on GAHD and HateCheck, (2) supporting annotators with a variety of methods not only increases their efficiency but also leads to more effective examples, and (3) GAHD is challenging, even for state-of-the-art LLMs and content moderation APIs.

Our results highlight the benefits of supporting annotators in finding adversarial examples. Future work could explore more annotator support strategies for DADC. Specifically, LLM-based augmentations (Bartolo et al., 2022), such as perturbations and counterfactuals (Qian et al., 2022; Sen et al., 2022, 2023) present a promising avenue.

Acknowledgements

We thank Rafael Mosquera, Juan Manuel Ciro Torres, the rest of the Dynabench team, and MLCommons for their support. The project received funding from the Swiss Federal Bureau of Communications (OFCOM), the University of Zurich Research Priority Program (project "URPP Digital Religion(s)"[9]), and the Linguistic Research Infrastructure of the University of Zurich. Paul Röttger is a member of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA).

Limitations

Annotator Demographics and Coverage GAHD aims to cover hate speech in the context of all three major German-speaking countries. However, we recruited our annotators only in one German-speaking country and instructed them to construct examples with protected groups and stereotypes from all three countries. Even though, when inspecting GAHD, we found evidence that the annotators succeeded in doing so, we acknowledge that the different countries are probably covered in different degrees.

Conversational Context We collected examples without conversational context. Especially examples that trick the target model via vagueness require imagining a context. Consequently, it is possible to envision a conversational context for some examples that would result in a different label.

Annotator Support Methods We observed that mixing multiple support methods lead to an overall more effective dataset. Since we were only able to evaluate three support methods, it remains open if the conclusion holds for other support methods.

[9] https://www.digitalreligions.uzh.ch/en.html

References

Izzat Alsmadi, Kashif Ahmad, Mahmoud Nazzal, Firoj Alam, Ala Al-Fuqaha, Abdallah Khreishah, and Abdulelah Algosaibi. 2021. Adversarial attacks and defenses for social network text processing applications: Techniques, challenges and future research directions. ArXiv: 2110.13980.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, and Heike Trautmann. 2021. RP-Mod & RP-Crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Rui Cao and Roy Ka-Wei Lee. 2020. HateGAN: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

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).

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Edita Grolman, Hodaya Binyamini, Asaf Shabtai, Yuval Elovici, Ikuya Morikawa, and Toshiya Shimizu. 2022. Hateversarial: Adversarial attack against hate speech detection algorithms on twitter. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 143–152, New York, NY, USA. Association for Computing Machinery.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

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.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 2023. Aporophobia: An overlooked type of toxic language targeting the poor. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, Toronto, Canada. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Pranath Reddy Kumbam, Sohaib Uddin Syed, Prashanth Thamminedi, Suhas Harish, Ian Perera, and Bonnie J. Dorr. 2023. Exploiting explainability to design adversarial attacks and evaluate attack resilience in hate-speech detection models.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. 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*, FIRE '20, page 29–32, New York, NY, USA. Association for Computing Machinery.

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*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium. Association for Computational Linguistics.

Melody Moh, Teng-Sheng Moh, and Brian Khieu. 2020. No "love" lost: Defending hate speech detection models against adversaries. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–6.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Rajvardhan Oak. 2019. Poster: Adversarial examples for hate speech classifiers. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 2621–2623, New York, NY, USA. Association for Computing Machinery.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. ArXiv: 2303.08774.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

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.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

Rupak Sarkar and Ashiqur R. KhudaBukhsh. 2021. Are chess discussions racist? an adversarial hate speech data set (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):15881–15882.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

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*.

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

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, Dublin, Ireland. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H. Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2017. Mastering the dungeon: Grounded language learning by mechanical turker descent. *CoRR*, abs/1711.07950.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598. Publisher: PeerJ Inc.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Ethical Considerations

Intellectual Property Rights Data created manually by the annotators does not violate intellectual property rights. The English adversarial hate speech dataset Vidgen et al. (2021) (used in R2) and the Leipzig Corpus Collection (used in R3) are both licensed under CC BY 4.0. According to this licensing, redistribution with proper attribution is considered fair use.

Intended Use This paper presents a dataset and methods intended to support the development of more robust and accurate hate speech detection models.

Potential Misuse: Spreading Hate Speech Actors that aim to spread hate speech while systematically evading content moderation could use this dataset as guidance. However, we believe that it is improbable that such actors identify critical model weaknesses that have not already been discussed and analyzed in public through this dataset. Further, by making this dataset publicly available, we support content moderation systems in making their models more robust against exactly the attacks that could be derived from this dataset.

Potential Misuse: Surveillance and Censorship Most research on methods for content moderation can be adapted and misused for surveillance and censorship. However, not working on content moderation has clear harmful consequences and leaves targets of hate, specifically marginalized minorities, vulnerable. As researchers who work on harmful language and NLP, we aim to conduct our research in a way that avoids facilitating its potential misuse.

B Initial Datasets

Table 5 contains the label distributions and additional details about our initial datasets.

We further preprocessed examples by removing excess whitespace, and by replacing user names (starting with "@") and URLs with placeholders.

The RP-Crowd dataset does not contain direct hate speech annotations, but rather scores for threats, insults, profanity, etc. We treated all comments with a sexism score or racism score higher than 2 as hate speech, and all other comments as not hate speech.

C Target Model Training Details

We list the hyperparameter used for training the target models in Table 6.

Initially, we experimented with higher learning rates of 5e-5 and 3e-5, but we found that 1e-5 leads to better performance. For all hyperparameters not listed in the table, we kept the default values of the trainer class from the huggingface transformers library (Wolf et al., 2020) (version 4.31.0). We always chose the checkpoint that performed best on the development set as the target model for the next round. For evaluation, we used sci-kit learn (Pedregosa et al., 2011).

Computation and Programming We ran all experiments on a cluster with eight NVIDIA GeForce RTX 3090 GPUs. Each GPU has 24 GB of RAM. Based on the fact that fine-tuning and evaluation of one target model on one GPU took approximately 40 minutes, we estimate that our experiments overall ran for ca. 60 GPU hours. We used GitHub Co-Pilot and ChatGPT for coding assistance.

D GAHD Examples

To give the reader an impression of typical texts found in GAHD, we provide an example for each GAHD topic from Figure 5 in Table 8. Further, as discussed in 3.5, we showcase vague or incomplete examples found in GAHD in Table 7.

E Evaluation of Large Language Models and APIs

Here, we provide additional details for the evaluation settings in Section 4.3:

Large Language Models We evaluated all LLMs with the same prompt containing a task description, a hate speech definition, and a response format. Figure 8 shows an example prompt. In the five-shot scenario, we added five randomly sampled entries, paired with their labels, from the training split. We sampled a new set of examples for each classification to average out the effects of specific examples in the prompt. For the GPT-models, we used JSON-mode[10] which guarantees that the models generate valid JSON. However, the LeoLM models were not able to respond consistently with valid JSON. We thus changed the response format for LeoLM to only one token: *TRUE* or *FALSE*.

[10]https://platform.openai.com/docs/guides/text-generation/json-mode

paper	name	train	dev	test	% hate	source
Demus et al. (2022)	DeTox	2,333	321	691	32.3	Twitter
Mandl et al. (2019)	HASOC 2019 Task 2	300	33	123	33.3	Twitter
Mandl et al. (2021)	HASOC 2020 Task 2	395	43	171	33.6	Twitter
Assenmacher et al. (2021)	RP-Crowd	2,130	304	608	32.6	newspaper
Röttger et al. (2022)	MHC (German)	-	-	3,645	70.0	synthetic

Table 5: Details of our initial datasets and of German HateCheck used in the evaluation.

parameter	value
epochs	5
learning rate	1e-5
batch size	8
gradient accumulation	4

Table 6: Hyperparameters of the target model.

We set the generation length to 1 ensuring that both tokens are present in the LeoLM vocabulary. If a LeoLM model responded with a different token we regenerated the response.

APIs The Perspective API does not provide categorical labels but scores between 0 and 1. We used the, by Google Jigsaw suggested, default threshold of 0.7[11] for mapping these scores to binary hate speech labels. The content moderation API from OpenAI provides scores as well as binary labels. We directly used the binary labels.

F Data Statement

Following Bender and Friedman (2018), we provide a data statement for GAHD.

F.1 CURATION RATIONALE

We had three motivations for building this dataset: (1) Exploring new methods for making DADC more efficient, (2) providing a resource to evaluate robustness for hate speech detection in German, (3) providing a resource to train more robust models for German hate speech detection. We further selected the English adversarial hate speech dataset (Vidgen et al., 2021), for being a large, high-quality, openly available, adversarial hate speech detection dataset. Finally, we selected the Leipzig Corpus Collection (Goldhahn et al., 2012) news corpus 2022 because it contains texts about current topics, is large enough for our purposes, and is permissively licensed.

[11]See: https://perspectiveapi.com/

F.2 LANGUAGE VARIETY

We instructed the annotators to create texts in standard German. Newspapers in German-speaking countries often require comment sections to be in standard German, but comments still sometimes contain expressions in a dialect. We account for this by specifically allowing annotators to sometimes use slurs from a dialect in an otherwise standard German sentence.

F.3 SPEAKER DEMOGRAPHICS

GAHD contains three separate speaker demographics: (1) The speaker demographics of the manually-created examples, are the same as the annotator demographics. We describe them in the next section. (2) For examples automatically translated from the dataset of (Vidgen et al., 2021) we refer to the speaker demographics of their data statement: https://aclanthology.org/2021.acl-long.132.pdf. (3) The speaker demographics of the newspaper data labeled in R3 are hard to characterize, as they contain sentences from a wide range of news websites. From that fact, we can assume that the speaker demographics mostly consist of German journalists. However, as described in Section 3.4, we found some sentences that rather look like newspaper comments sentences out of a newspaper article.

F.4 ANNOTATOR DEMOGRAPHICS

Section 2.4 already contains information on annotator demographics. Here, we repeat the information and provide additional details: We distributed the annotation load between as many annotators as possible while keeping the administrative overhead manageable and in line with university requirements. This led to the recruitment of seven annotators at our university. Three of the students were female (43%), three were male (43%) and one was non-binary (14%). Three annotators had a high school diploma and were currently pursu-

Example	English translation
Das hat Mama Maye der „	That has Mum Maye the “
Sie Weihnachten Monat.	She Christmas month.
Gurten gegen internationale "Auswahl"	Gurten against international “selection”

Table 7: A list of incomplete, grammatically incorrect, or vague, examples found in GAHD, which we chose to leave in the dataset as they fall not under the definition of hate speech and are thus valid instances of not-hate speech.

ing a bachelor's degree (43%), three had a bachelor's degree and were pursuing a master's degree (43%), and one annotator had a PhD and worked as a postdoc (14%). Five were native German speakers (71%) and two were highly proficient but non-native speakers (29%). Six annotators were in the age range of 18-29 (86%), and one annotator was in the age range of 30-39 (14%). For the last round, we recruited two additional annotators who worked at the university. Both were male, had a master's degree, were native German speakers, and in the age ranges of 30 to 39, and 40 to 49. The lead author took the role of expert annotator. He is a male, native German speaker with a master's degree and in the age range of 30 to 39.

All annotators had basic or advanced knowledge of computational linguistics. Three annotators already had knowledge about or experience with hate speech detection, which they gained through coursework or student projects.

We paid the annotators over 30 CHF per hour, according to university guidelines. We spread the DADC rounds over four months, with a data collection window of two to four weeks per round. This gave the annotators the freedom to schedule their working hours in a way that fits their other duties. After each round, the annotators reported how many hours they had worked.

Before the first round, we held a 1.5-hour presentation and discussion session where we gave the annotators an overview of the project, in-person instructions, and provided a space to discuss the definition of hate speech. The annotators then worked remotely. We analyzed the submitted examples and annotations after each round. If necessary, we provided feedback and further instructions via online meetings and a group chat.

F.5 SPEECH SITUATION

The data creation and labeling took place between July 2023 and November 2023.

F.6 TEXT CHARACTERISTICS

We describe the label distribution and general topics present in GAHD in Section 3.6.

topic	example (German // English)	R
1: the COVID-19 virus and its impact	Großartige Neuigkeiten ! Schwarze Amerikaner verzeichnen landesweit nach wie vor die höchsten tatsächlichen COVID-19-Sterblichkeitsraten, die mindestens doppelt so hoch sind wie die Rate für Weiße und Asiaten, die die niedrigsten tatsächlichen Raten aufweisen. *Great news ! Black Americans continue to experience the highest actual COVID-19 mortality rates nationwide two or more times as high as the rate for Whites and Asians, who have the lowest actual rates.*	2
2: texts discussing Turkish people and culture, some with negative stereotypes	In Kramerhof in Mecklenburg-Vorpommern wurde ein Wehrdienstleistender entlassen, der äußerte, dass er mit einem „Scheiß Türken“ auf der Stube sei und gerne sehen würde, „wie er den Bordstein frisst“. *In Kramerhof in Mecklenburg-Western Pomerania, a military serviceman was discharged because he said he was in the room with a “shitty Turk” and would like to see him “eat the curb.”*	3
3: the relationship between Ukraine and Russia	Egal wie, die Anexion der Krim durch Putins Russland, hat keine Ursache in der Ukraine, sondern in der stalinistischen "Blut und Boden" Naziideologie in Moskau. *No matter how, the anexation of Crimea by Putin’s Russia has no cause in Ukraine, but rather in the Stalinist "blood and soil" Nazi ideology in Moscow.*	3
4: derogatory language towards people from Pakistan	Die Art und Weise, wie Bangladescher Pakistaner hassen, ist brillant!! Sie haben völlig Recht, sie wissen es besser als jeder andere, es kommt so viel Wahrheit über den pakistanischen Bullshit ans Licht *the way that bangladeshis hate Pakistanis is brilliant!! They’ve totally got it right, they know it better than anyone, so much truth comin out from them about the pakistanis bullshit*	2
5: stereotypes and generalizations about African people	An die, die sagen, die Afrikaner seien unzivilisert: Überlegt mal, welchem Kontinent wir alle entstammen.. *To those who say that Africans are uncivilized: Think about which continent we all come from.*	1
6: Negative stereotypes about people from the former Yugoslavia	Ich schicke mein Kind doch nicht in eine Schule mit Jugos, die sollte man in eine andere Schule stecken *I’m not sending my child to a school with Yugos, they should be put in another school*	4

7: the integration and treatment of disabled individuals	Dass Behinderte aus der Gesellschaft entfernt werden müssen, können auch nur die Rechten behaupten. // *Only the Right would claim that disabled people must be removed from society.*	4
8: immigration and national identity in Germany	In diesem Land geht es immer schlechter, da hier immer mehr Asiaten und Schwarze leben. Es muss etwas getan werden, um sie zu stoppen *This country is getting worse because more and more Asians and Blacks are living here. Something must be done to stop them.*	2
9: migration policies and their impact on public services	Migranten haben große Familien, was ihnen viele soziale Vorteile bietet *Migrants have large families, which provides them with many social advantages*	1
10: urbanization and gentrification in various cities	Leute die weniger verdienen werden zunehmend aus den Städten verdrängt. Das muss endlich aufhören! *People who earn less are increasingly being displaced from cities. This must stop!*	1
11: negative attitudes towards refugees and their impact on society	Die Grenzwächter sollten Maschinengewehre haben damit die Flüchtlinge draussen bleiben *Border guards should have machine guns so that refugees stay outside.*	1
12: politicians, police, and trust in people with Polish roots	Politiker sind der schlimmste Mist überhaupt *Politicians are the worst shit ever*	2
13: football teams and players	Stattdessen sollte man sie mit den positiven Dingen locken, die dann passieren würden, wenn sie dem besseren Club beitreten. *Instead, they should be enticed with the positive things that would happen if they joined the better club.*	3
14: discussing Islam and Muslims in a neutral manner	Viele andere europäische Länder weigern sich auch, Flüchtlinge zu nehmen, weil sie lieber keine Musliminnen aufnehmen wollen. *Many other European countries also refuse to take refugees because they prefer not to accept Muslims.*	4
15: various topics and perspectives	STELLEN SIE SICH EINE WELT OHNE DEICHE VOR. Jetzt machen wir es möglich!! *IMAGINE A WORLD WITHOUT DAMS. Now we're making it possible!!*	2
16: offensive language and racial slurs	Ich liebe meine Rasse, ich bin so froh, dass ich nicht als N***r geboren wurde *I love my race, I'm so glad I wasn't born a n***r*	2

17: anti-Semitic hate speech	Mit diesem Tweet rufe ich zur Jagd aller Juden aus // *With this tweet, I call for the hunting of all Jews.*	1
18: gender roles and women's rights	Und sie sagen, man könne Frauen trauen. lol was? *And they say you can trust women. lol what?*	4
19: the experiences and treatment of black people	Polizeigewalt gegen Schwarze sei ein gewaltiges Problem. Verletzte, Tote, zerstörte Leben... Egal, sind ja eh nur Schwarze. // *Police violence against black people is a huge problem. Injured, deaths, destroyed lives... Doesn't matter, they are just black anyway.*	1
20: mental health and psychological behaviors of people	Die damaligen Einwohner Kanaans hätten die psychoaktive Droge offenbar als «Beigabe für die Toten» verwendet, hieß es. *The former inhabitants of Canaan apparently used the psychoactive drug as an "addition for the dead," it was said.*	3
21: gender issues and LGBTQ+ rights	Trans- und Intersexuelle müssen in der Lage sein, ihre Art mit zivilrechtlichem Status erleichtert zu ändern. . . von Menschen zu Sub-Menschen *Trans and intersex individuals must be able to change their gender with facilitated legal status... from humans to sub-humans.*	4

Table 8: An example for each topic in GAHD, as identified in Section 3.6. Hate speech examples have red borders and non-hate speech examples green borders.

Prompt

System Message

Your task is to predict whether a given piece of text in German is hate speech or not. Use the following definition for hate speech: Hate speech is abusive, discriminatory, derogatory or dehumanizing speech targeting a protected group or a person for being member of such a group. Protected groups in the German-speaking context are marginalized groups defined via the following characteristics: race, religion, ethnicity, national origin, gender, sexual orientation, disability. Majority groups or groups defined via other characteristics are not protected groups. Respond in JSON format with the key "label" and the value 1 (int) for hate speech or 0 (int) for not-hate speech.

Input-Output Examples for a Five-Shot Scenario

Here are five examples of text and the corresponding output:
Example 1: "Sehr erfrischend ist auch Umgang mit Behinderten im ganzen Land. Sie sind nirgends zu sehen."
Output 1: "{'label': 1}"
Example 2: "Ich hassse diesen Sch**sser, am liebsten würde ich ihn abstechen!!"
Output 2: "{'label': 0}"
Example 3: "Frauen gehören sich selbst, nicht der Küche."
Output 3: "{'label': 0}"
Example 4: "Dunkelhäutige Menschen gibt es überall, ok? Hör auf mit deiner Xenophobie!"
Output 4: "{'label': 0}"
Example 5: "Der Türke ist Fussballprofi und Restaurantchef. Und er ist sehr dumm, gut das kann man ableiten aus seiner Herkunft."
Output 5: "{'label': 1}"

User Message

New text: "Ich glaube nicht, dass Schwarze dumm sind. Sie sind sehr klug. Machen nämlich den ganzen Tag nichts und leben von der Sozialhilfe, die WIR bezahlen."
Output:

Model Output

{"label": 1}

Figure 8: Five-shot prompt for GPT models.