Building a Benchmark Dataset for Robust German Hate Speech Detection

Final Report [April 16, 2024]

Gerold Schneider, Janis Goldzycher
Department of Computational Linguistics
University of Zurich
Andreasstrasse 15, 8050 Zurich

This report is based on the paper "Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset" authored by Janis Goldzycher, Paul Röttger, and Gerold Schneider. It will be published and presented at the Annual Conference of the North American Chapter of the Association for Computational Linguistics 2024 (NAACL 2024). We write this report as an extension of the paper to provide additional background information and contextualization.

Executive Summary

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 project, 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`.

1 The Context of our Research: Hate Speech and its Detection

1.1 Motivation

Hate speech is a serious problem in our society, in particular on social media. In the last few years hate speech has increased in frequency. After Elon Musk's take-over of Twitter and re-branding it to X, racist posts have increased in frequency [1]. Due to changes in algorithms, Twitter/X users are also more likely to be exposed to hate speech[2]. The social media index GLAAD observed a decrease in user safety scores for the LGBTQ+ communities for the second year in a row for all leading social media platforms, Facebook, Instagram, TikTok, YouTube and Twitter [3].

At an intuitive level, hate speech is unpleasant and everyone agrees that it hurts the attacked groups or individuals. Even the attackers would agree, it is even a main motivation why they use hate speech. Up to a point, attackers may deliberately risk or even encourage the violation of the human rights of liberty and security, and even risk the right to life. The unobtrusive modal verb *may* in the above sentence is also part of the problem: a unpremidated bout of

[1] `https://www.ohchr.org/en/statements/2023/01/freedom-speech-not-freedom-spread-racial-hatred-social-media-`
[2] `https://www.washingtonpost.com/technology/2023/03/30/elon-musk-twitter-hate-speech/`
[3] `https://glaad.org/publications/social-media-safety-index-2023/`

anger may equally hurt of frighten an attacked person as a well-planned death threat if the context is unknown.

At the same time, censoring hate speech poses a serious problem: it violates the human right of free speech, freedom of opinion and expression. Therefore, hate speech leads to a clash of human rights and a bias towards excluding some voices.

1.1.1 A clash of human rights

There is no unanimity of what the basic human rights are. Let us consider an example.

According to the learning platform Vaia[4] a ten-item list of fundamental human rights comprises:

1. Right to life: Every person has the right to live and not be deprived of life unlawfully.
2. Freedom from torture: No person should be subject to torture or cruel, inhuman, or degrading treatment.
3. Right to liberty and security: Everyone has the right to be free from arbitrary arrest or detention.
4. Freedom of thought, conscience, and religion: All individuals have the right to hold and practice their beliefs freely.
5. Freedom of opinion and expression: People have the right to hold and share opinions and ideas without interference or censorship.
6. Right to work and education: Everyone has the right to work in fair and safe conditions and to receive an education.
7. Right to privacy: All individuals have the right to privacy in their personal, family, home, and correspondence lives.
8. Right to participate in government: Every person has the right to take part in their country's political affairs and exercise their right to vote.
9. Freedom of movement: People have the right to move freely within their country and to leave and return to it.
10. Right to equality before the law: All individuals are entitled to equal protection of the law without discrimination.

Censoring hate speech violates rights 4 and mainly 5, while unchecked hate speech may violate rights 4 and jeopardize rights 1 and 2 and possibly 3. Right 4 is particularly interesting, as it may be seen as comprising a clash in itself: freedom of thought entails the right to hate, while freedom of religion entails the right to practice any religion without fear. A crucial difference between rights 4 and 5 is, though, that while thoughts need to be free (right 4), uttering them without check (right 5) may lead to hate speech speech which puts rights 1 to 3 in danger.

Defenders of unlimited free speech typically argue that verbal expressions cannot lead to bodily harm, thus rights 1 to 3 are guaranteed, but this is not true for two reasons: first, clearly stated intentions encourage real-world violence threatening these rights, and secondly, because attacked groups and people may psychologically suffer from degrading treatment in verbal form. Also, further rights can be affected: freedom of movement (right 9) can be subjectively affected, for example, if some social groups subjectively feel that they are systematically discriminated. Psychological harm is very difficult to measure and typically used by both attacked groups and the utterers of hate speech, both feeling threatened, sometimes both due to the above first reason

[4] `https://www.hellovaia.com/explanations/law/human-rights-law/fundamental-human-rights/`

— clearly stated intentions encourage real-world violence. We come back to this point in the next subsection.

In a situation of such a clash, both radical answers, either prioritizing free speech unconditionally and thus allowing all hate speech utterances, or filtering everything that may hurt personal feelings, are untenable. There is no unanimous definition of what hate speech is, this is an active research question, see e.g. Hietanen and Eddebo (2023), and it is generally agreed that vulnerable groups need to be protected Waldron (2012).

Waldron (2012) also shows how for U.S. constitutionalists regulation of hate speech may be seen as violating the First Amendment of the U.S. constitution. Particularly in conservative politics the regulation of hate speech is seen as a political issue. The viewpoints of Donald Trump and Elon Musk have brought to worldwide attention that this clash exists and that there are voices that want to prioritize freedom of speech over other human rights, and the view that even the definition of hate speech is a purely political question. In politics and economy, both liberalist fully free-market and fully protected communist ideologies are seen as failed by most theorists. In a similar vein, it is likely that fully unchecked hate speech will lead to disaster equally as filtering all negative comments. But we need tools that detect hate speech allowing social media platforms, regulators and the government to take appropriate measures. Our contribution is to provide such tools.

1.1.2 "Die Gedanken sind frei" — Free thoughts on free thought

While elucidating the clash between human rights in the previous subsection we have already hinted at the next clash: while thoughts must be free and cannot be controlled, not even in the most oppressive dictatorship, the correlation of thought and action is a classic in philosophy, ranging from commonsense utterances that action speaks louder than words and one should judge people (both in moral and legal terms) not on their thoughts but on their actions, to the reminder that there is a connection, for instance in the words of Nobel prize winner Maria Ressa, also quoted in Hietanen and Eddebo (2023):

> "Online violence does not stay online. Online violence leads to real world violence."
> —Maria Ressa, Recipient of the Nobel Peace Prize (SVT, 2021, 1:04:03)

While there is unanimity that the condemnation of violence is a cornerstone of civilized society, there is disagreement (1) on whether verbal expressions as such constitute harm and (2) on whether hate speech incites violence in the real world.

Point (1) has several aspects: while is clear that psychological harm may be as hurtful as physical harm, it is harder to measure. Theoretical discussions do not necessarily see hate speech as harm per se. For example Barendt (2019), who observed that "In Jeremy Waldron's book, The Harm in Hate Speech [Waldron (2012)], it is not always clear whether he argues that hate speech causes harm or whether it constitutes harm." His conclusions are as follows:

> If the right to free speech is taken seriously, strong arguments must be advanced to justify its restriction and evidence adduced to establish a link between hate speech and the harm it is alleged to cause. It would be unreasonable to expect this evidence to be provided in Waldron's book, which is concerned with putting forward general arguments of political principles. The best interpretation of his argument is that it is legitimate to ban hate speech because it has harmful tendencies to endanger social cohesion and injure the dignity of targeted groups. That is the weak form of consequentialist argument: hate speech may be banned because of a general apprehension of its effects, not because there is evidence that it really does cause substantial harm, whether to social order or its victims (Section 3). This argument leaves much to the judgment of government when it is appropriate to intervene; for that reason alone it is unattractive to advocates of the free speech principle who are suspicious of government regulation of freedom of speech (Schauer 1982, 85-6).

Point (2) is easier to argue for, and also Barendt (2019) links to it: if hate speech leads to physical harm, it needs to be detected and censored, particularly if vulnerable minorities are affected. Also, defendants of direct democracy sometimes forget that the touchstone of democracy is not only the rule of majority but equally the protection of minorities. The United Nations are unequivocal in stressing that hate speech is often the precursor to real violence, history has taught us in many cases ranging from the holocaust to the Srebrenica genocide in Bosnia and Herzegovina[5]. There is mounting evidence that online hate can turn into real-life violence[6]. Williams et al. (2019) investigate correlations between police crime and Twitter data to show that there is a positive correlation between social media hate and real-world crime. The authors conclude: "This research shows that online hate victimization is part of a wider process of harm that can begin on social media and then migrate to the physical world."

Correlation studies do not, strictly speaking, measure cause and effect, which may be adduced as an argument against the findings of Williams et al. (2019). Psychological studies now also support the intuition that there is a cause-and-effect relation, with hate speech possibly causing mental differences: Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain Pluta et al. (2023). From a utilitarian perspective, it is also important to mention that calls for hatred and violence are pointless if they are not meant seriously in the vast majority of cases. If all authors of hate speech were convinced that there is a complete disjoint between actions and words there would be very little hate speech.

1.1.3 Why is hate speech harmful, and its detection important for society?

After zooming in on the clashes of human rights and the freedom of speech, let us broaden the perspective again and remind ourselves of a dozen or so reasons why hate speech is harmful and may have a corrosive effect on society, and thus why tools such as ours are needed. This list is not encompassing.

- it affects participation and inclusion
- creates fear and anxiety among the targetted groups
- children are at particular risk
- it divides and polarizes society
- it is often not based on facts
- does not contribute to a solution to possibly real problems
- it may violate human rights
- it is often a precursor to real violence
- it meets no immediate resistance in the anonymous space of the internet
- if unanswered, not met with resistance such as counterspeech may lead to radicalization (echo chambers)
- it is often also harmful for the attackers: fit of anger may lead to exclusion, loss of reputation or even job
- it is a real threat to democracy

1.1.4 Why can hate speech not be filtered manually?

- sheer mass: 1-4% of messages contain hate speech
- psychologically affects people if constantly exposed to hate speech
- very tedious and repetitive
- platforms are under financial constraints
- EU regulations demand reliable detection (Musk VS EU)
- no clear definition of hate speech: https://unesdoc.unesco.org/ark:/48223/pf0000379177

[5] https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

[6] https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/

Frequency for	christ*	jew*	muslim*
doc1	6	2	1
doc2	2	3	6
doc3	1	4	5

Table 1: Document-term matrix for the example in Figure 1

1.1.5 Why is automated hate speech detection difficult?

If even a clear definition of hate speech is difficult, its detection will be all the more difficult. Hate speech, or in general expressions with a similar meaning, may be expressed in a multitude of ways. At the level of words, linguists speak of synonymity and ambiguity.

So-called synonyms express very similar meanings. For instance, *astronaut* and *cosmonaut* are synonyms, or *hate* and *despise*, or *kill* and *execute*. In order to cover all expressions potentially containing hate speech, one would need a very long list of words, and a large dictionary. These examples of synonyms also show that there are hardly any full synonyms: *astronaut* points to a U.S. or European setting, will *cosmonaut* refers to Russian space programs, with all the political and military implications. *kill* refers to any form of taking life, while *execute* has more likely a legal setting. These subtle differences may have an effect on hate speech status: while *I think all Jews should be killed* is clearly hate speech, *I think all terrorists should be executed* is clearly not – discussions on the death penalty need to be possible in democracies.

While the word *execute* on the one hand has a narrow meaning when it refers to taking life, it also has further meanings that are very different — it is highly ambiguous. Think of the contrast between *I think all police orders should be executed* and *I think all police staff should be executed*. Ambiguity is a main reason for using a large dictionary of words, so-called dictionary-based approaches do not perform very well. In the example *execute* many, probably the majority of utterances do not contain hate speech. In *I think all police orders should be executed* the fact that police orders are not alive (linguists use the term *animate*) triggers the correct reading of *execute* and also illustrates that the status of an utterance hate speech or not depends on whether the hate speech target is animate and member of a protected group.

These examples show that we need to know more than individual words that we use as a filter. We minimally need to use **words in combination and their interaction**, and we need to know **which words are similar**.

1.2 Recent Developments in Text Technology

Text Technology, disciplines like Computational Linguistics and applications like media content analysis offer methods addressing these requirements. A relatively simple method to use words in combination is bag-of-words classification, and word similarities can be computed by the various methods that are called word embeddings.

https://fra.europa.eu/en/publication/2023/online-content-moderation

1.2.1 Bag-of-words Classification

Decisions that are based on all words found in a document are more reliable than those depending on a small list of dictionary words. Document classification systematically considers all words, although rare words and function words (such as articles and prepositions, they are also called stopwords) are excluded. How should all words be included in such a comparison in an efficient way?

In order to efficiently compare documents, be it to assign a document to a class (e.g. hate or not-hate, or religion, or political affiliation etc.), efficient representations allow the comparison of documents. One of the most widely used approaches is so-called vector-space representation. In them, every word type, or at least every keyword used in the document collection, is represented

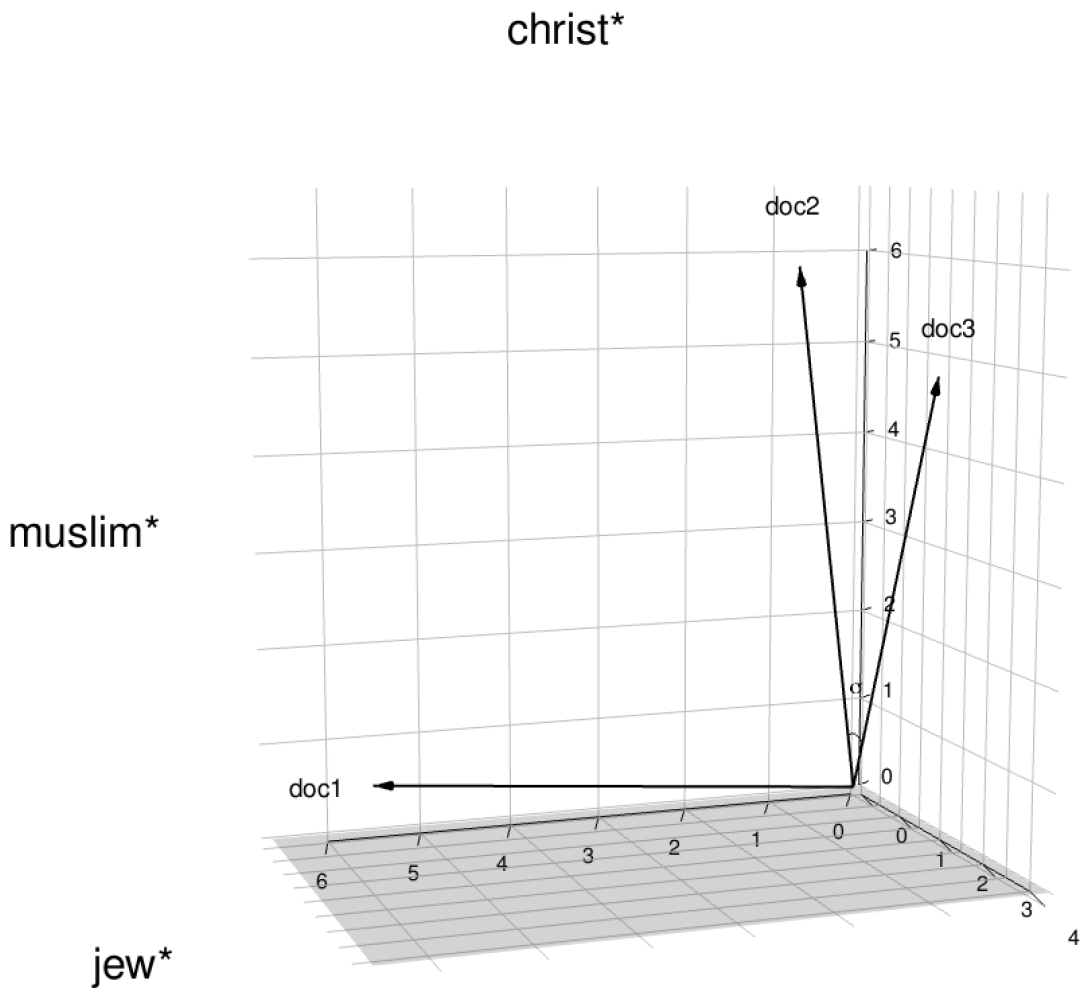

Figure 1: Vector space representation of the document-term matrix in Table 1

as a dimension. In our example, we will show an example using stemmed words, for instance *christ** covering *Christ, Christian, christianity*. The value of the dimension corresponds to the frequency (or a keyword value) of the word in the given document, such as a post on social media. Let us consider an example suitable for religion classification. For purposes of the simplicity of illustration, we assume three documents in which the only words that we consider are *christ**, *jew** and *muslim**. The frequency of these terms in three documents (doc 1 to doc 3) is given in Table 1, the vector-space representation in Figure 1. Table 1 is a so-called document-term matrix.

Figure 1 shows, for example, that doc2 and doc3 are more similar than doc1. This can be seen as the vectors pointing into a similar direction, and it can be systematically assessed by measuring the angle σ between doc2 and doc3. The cosine of the angle delivers a value between 0 and 1 expressing the similarity of the two documents, 1 for identity (cos 0°) and 0 for maximum difference (cos 90°).

Vector representations are used for many tasks. In document classification, a new document that needs to be classified can be given the same class as one of the most similar annotated documents. The example of Figure 1 and Table 1 uses only 3 words, for the simple reason that we cannot imagine coordination systems with more than three dimensions. In a typical setting, thousands of words are included. Although we can no longer imagine this vector space, the cosine measure works in the same fashion. We stated above that one needs to use **words in**

combination and their interaction. Document classification respects the combination of words in the sense that it includes all words in a given document. The interaction model is a radically simple one, though. It is only counted how often a word occurs in a given document. The sequence of words and their position in the document or in the sentence is not taken into consideration. That is why this method is often called a bag-of-words model. In order to include a minimal notion of word order, the model is often extended to include frequent sequences of two words (bigram model) or three words (trigram model). Longer sequences are typically not used, as most longer sequences have very low frequencies.

The way in which the thousands of features are combined to make a class prediction can be done with several algorithms, which are also well-known from other domains of social and natural science, ranging from Naïve Bayes (in which all features have the same weight) to logistic regression (in which each feature learns its optimal weight) to support-vector machines (SVM) which also manage to capture non-linear relations. Instead of single classifiers, multiple classifiers can be arranged into an ensemble system or a neural network in which each node can be seen as a logistic regression classifier.

1.2.2 Word Embedding

We stated that a further requirement for successful hate speech detection, in fact, any automated content analysis is to obtain knowledge on **which words are similar**. This can be extracted from large collections of texts, by considering the typical contexts of every word. Similar words tend to occur in similar contexts, as human language is inherently redundant. The Firthian hypotheses, summarized by "You shall know a word by the company it keeps" Firth (1957) allows one to detect similar words from the sums of their contexts and can thus add semantic knowledge to language models.

Vector models can also be used to calculate semantic similarity. Instead of a document-term matrix as in Figure 1, we build a term-term matrix in which words that co-occur within a context window of for example 10 or 20 words. Sahlgren (2006) shows that while a very narrow context such as word adjacency delivers linguistic collocations, i.e. relations at the syntagmatic level, broader context windows, such as 10 words before and after, deliver semantic relations and associations, i.e. relations at the paradigmatic level. This insight is directly exploited by the hugely successful research paradigm of distributional semantics Baroni and Lenci (2010), which aims to detect synonyms, antonyms and hyponyms of words.

The same cosine metric as in document classification then delivers semantic similarity of words. As the term-term matrices and resulting vector spaces are very high-dimensional, often several thousand of dimensions, and very sparse, which means that most cells have a value of zero, various techniques of dimensionality reduction are used. They use well-known and efficient vector calculations. Frequently used methods are singular value decomposition (SVD, Deerwester et al. 1990) and principal component analysis (PCA, Pearson 1901). What all dimensionality reduction methods have in common is that they aggregate similar features.

In addition to vector-space models, models predicting similarity based on neural networks are frequently used. This approach is called word embedding, and it performs slightly better than vector-space models Baroni et al. (2014).

1.2.3 Supervised, unsupervised, and self-supervised learning

While document classification needs texts that are annotated for the classes that the algorithm should be able to detect for new documents, word Embeddings are learned purely from the texts.

Document classification, and more generally the algorithms that it employs, for instance, logistic regression, are typical instances of supervised learning, while word embedding, and more generally all clustering approaches, are instances of unsupervised learning.

More recently, an approach called self-supervised learning has become very influential, as it is the background of Large Language Models such as BERT and GPT, the latter is the base

for the famous ChatGPT tool. Supervised learning approaches typically perform better than unsupervised approaches, but annotating data is very labour-intensive. It is usually not possible to annotate millions of documents. Unsupervised learning has the advantage that it can profit from the almost unrestricted amounts of data available today, such as complete web scrapes and Wikipedia dumps. Self-supervised learning takes these enormous amounts of textual data and makes class predictions that are readily available, although they may seem to be very far away from the prediction that is required for a given annotation task. Self-supervised learning predicts the next word (this is why they are also called generative models, as they can directly generate text based on an initial sequence, for instance, a sentence), some models also predict missing words (gap filling, like in a cloze test) or the full sentence. BERT models focus on predicting missing words: every 15th word is masked and the training process learns to predict it as accurately as possible. For this reason, these models are sometimes also called masked language models. Although self-supervised models are basically trained for the "wrong" task unless you want to predict word sequences, their world knowledge is impressive. They have seen more text than an experienced human in their entire life. Due to this, they typically only need little adaptation to be tuned to a specific task, such as question answering, natural language inference, text summarisation or hate speech detection.

1.2.4 Large language models

The models that emerge are several orders of magnitude larger than the largest supervised models. The number of features (often also called parameters) used for document classification or Distributional Semantics with vector models is roughly 10^4. BERT models and the first GPT model (GPT-1) have about $10^8 = 100$ million parameters. BERT base has 110 million, BERT large 345 million, while GPT-1 has 117 million parameters. If a logistic regression model takes a minute to train, a corresponding BERT or GPT-1 model would take approximately $10,000$ minutes, which is seven days. GPT-2 has 1.5 billion parameters ($1.5x1,000$ millions$= 10^9$), while GPT-3 even has 175 billion parameters ($1.75x100,000$ millions $= 10^11$). If training time for a logarithmic model with 10^4 parameters is a minute, we would have to face training times between a month and ten years. Also, the neural network architectures needed are also much more complex. In a typical feed-forward neural network of 10 layers x 10 nodes we have 100 $= 10^2$ nodes, and each node in a neural can be thought of as a separate logistic regression. The architecture of a transformer is more complex (Vaswani et al. 2017), in particular the arrangement of the connections. The number of layers varies, for GPT-1 there are 768, GPT-2 has 1600, and GPT-3 12288 layers. The arising complexity also affects the complexity of the calculations. As the models are so complex, and often perform as well as document classification, but on a large variety of tasks, users no longer train then from scratch, which would also be unecological. Training a GPT-3 model from scratch uses as much energy as a thousand US households per year.

1.2.5 Neural Networks and Transformers

We mentioned each node in a neural can be thought of as a separate logistic regression. In addition to the logit function known from logistic regression, other activation functions can also be used to trigger a node or "neuron" to fire or not[7].

Classical feed-forward networks (see Figure 2 arrange the nodes in a grid of several layers, each layer containing several nodes, and every node is connected to each node in the subsequent layer. The number of layers defines how "deep" the deep neural network is. These networks have been used successfully for many tasks, also in computational linguistics. While they often performed better than e.g. logistic regression classifiers, they are not optimal for important dependencies that stretch across long sequences, such as pronoun resolution (the antecedent can

[7]See `https://en.wikipedia.org/wiki/Activation_function#Comparison_of_activation_functions` for an overview of activation functions

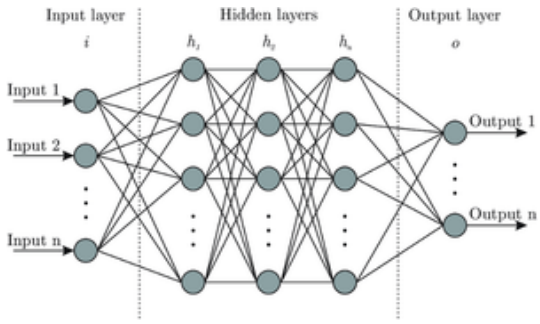

Figure 2: Illustration of a feed-forward neural network.
Source: https://miro.medium.com/v2/0*nDmq2u6JNCjZCd-A.png

be far away) or syntactic relations. Language is essentially sequential, and transformers have been designed to be able to respect the sequential character of language, in which some words that have passed considerably before the current word are important, while others are not. For a while, LSTM (long short-term memory) networks were used for this, but transformers Vaswani et al. (2023) systematically perform better. The architecture of a transformer network is shown in Figure 3.

Like in document classification and distributional semantics, the input text is tokenized and then converted into a vector representation[8]. This vector space represents every token of the same type in the same way. Then, the layers of the network are alternatingly feed-forward and attention layers. In the attention layers, each token is contextualized within the scope of the context window — which is usually the entire sentence — by means of the so-called attention mechanism. The attention mechanism amplifies the weights of (the few) important tokens and decreases the weight of all others. The aim of the attention mechanism is to simulate cognitive attention, the ability to attend to what is crucial for a given task and recognize all other data as irrelevant. The effect of this architecture is that it allows the model to access any preceding point along the sequence directly, instead of only indirectly via intermediate layers.

1.2.6 Pre-training and fine-tuning

The large pre-trained models can be used directly for many tasks, without any adaptation, a so-called zero-shot approach. Alternatively, they may be adapted to a task with a small number of additional training instances. These approaches are called few-shot. Usually, only the weights of the last few layers of the network are adapted based on the annotated, task-dependent material or a further task-specific layer is added. The main advantage of fine-tuning is that far fewer training instances are needed than when training a model from scratch. Transformer-based LLMs have such detailed world knowledge that fine-tuning only needs to specify the particular task, for example, question answering, summarization, natural language inference, stance detection, hate speech detection, language level, etc.

1.3 Shortcomings and underlying reasons

A common criticism of deep neural networks is that they are "black-boxes", methods that are too complex to understand what happens in detail. Accordingly, it is hard to anticipate in which cases these models tend to fail. Neural models including transformers are typically very reliable, but sometimes they produce arbitrary, seemingly absurd results in sparse data situations, so-called hallucinations. Traditional model evaluation, i.e. computing the accuracy or F1-score over an entire test set, does not help in this situation, since it only measures the overall performance – not the specific strengths and weaknesses of a given model. As a solution to this lack of

[8]This paragraph is partly a summary of the Wikipedia entry `https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)`

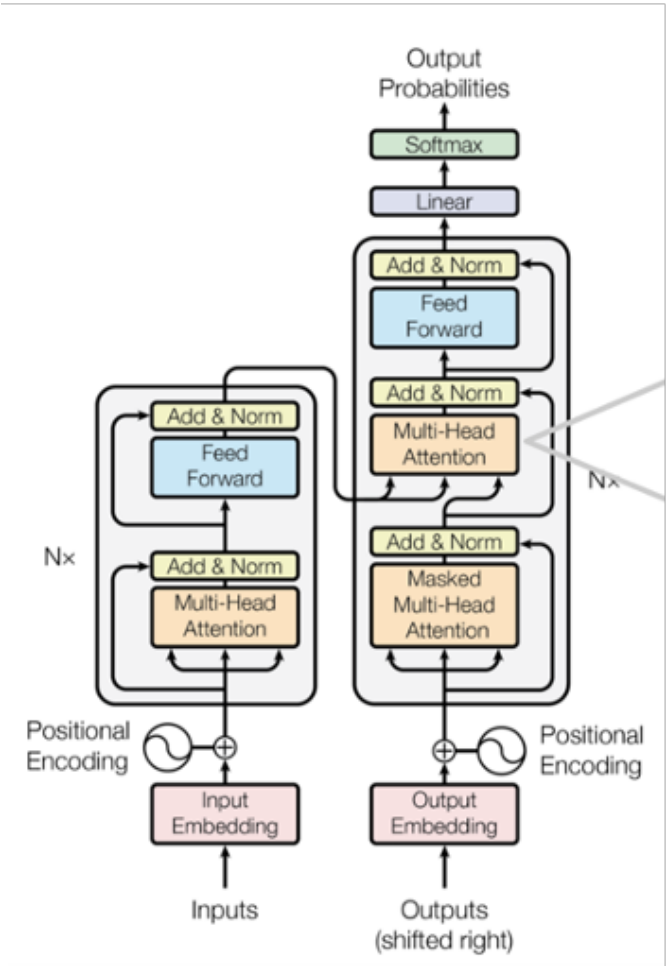

Figure 3: The architecture of a transformer. From Vaswani et al. (2023)

insight where models fail behavioral testing has been suggested Ribeiro et al. (2020). This is the motivation for our project. In addition, adding difficult and rare cases systematically, as we are doing, also reduces the risk of hallucination.

In addition to the algorithmic shortcomings, to which we alluded above, and which we evaluate in detail the results section, we are aware that our methods have many further shortcomings. For instance, it can be argued that detecting hate speech only finds the symptoms but does not address the underlying questions: why do some people feel so offended, marginalized, and threatened by society that they see no other way but to resort to uttering hate speech? Will people who feel patronized by the state ("Wutbürger") really feel less patronized by AI?

As partial answers, we could state that recognizing hate speech is a first step. Deleting offenders' posts or banning users if abuse persists at least protects the potential victims, the targetted groups. Counter-speech and talking to the identified offenders about their situation is a second step that one can hopefully take.

2 Project 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).

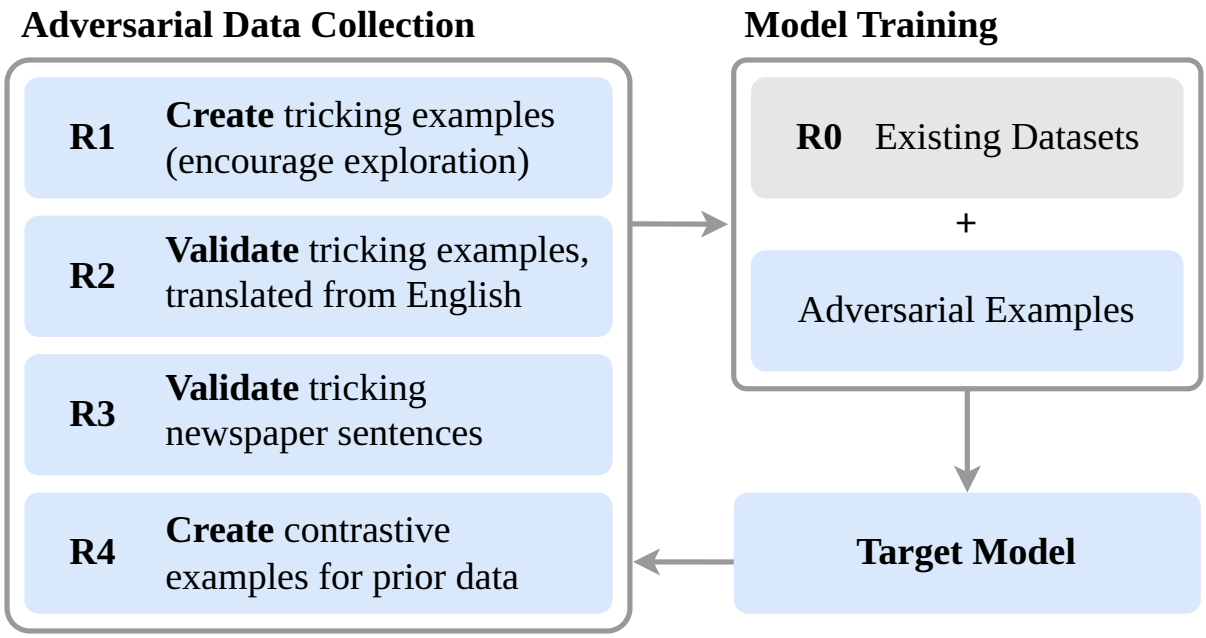

Figure 4: We use four rounds of **dynamic adversarial data collection** (Kiela et al., 2021) to improve a German hate speech classifier. We start with a target model trained on existing datasets. Then, in each round (R1-R4), annotators try to trick the target model using a different method. After each round, we train a new target model including the new adversarial examples.

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). 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), for example, 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 report, we give an introduction to hate speech detection and then introduce GAHD, a new **G**erman **A**dversarial **H**ate speech **D**ataset, collected with four rounds of DADC. However, 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 4 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 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 labels.

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.

3 Background

3.1 Hate Speech Detection

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 that models rely 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).

3.2 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).

4 Annotation

4.1 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.

4.2 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).

4.3 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 providing an elaborate hate speech definition the instructions contain three main points: (1) They specifically emphasize the culture-dependence of hate speech, 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 the dataset 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.

4.4 Annotator Demographics

To support diverse model-tricking strategies we distributed the annotation load between as many people as possible., constrained by our budget and university requirements. 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.

5 Dynamic Adversarial Data Collection

5.1 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).[9] We chose this model because it is both strong and light-weight, so that annotators receive fast feedback 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

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

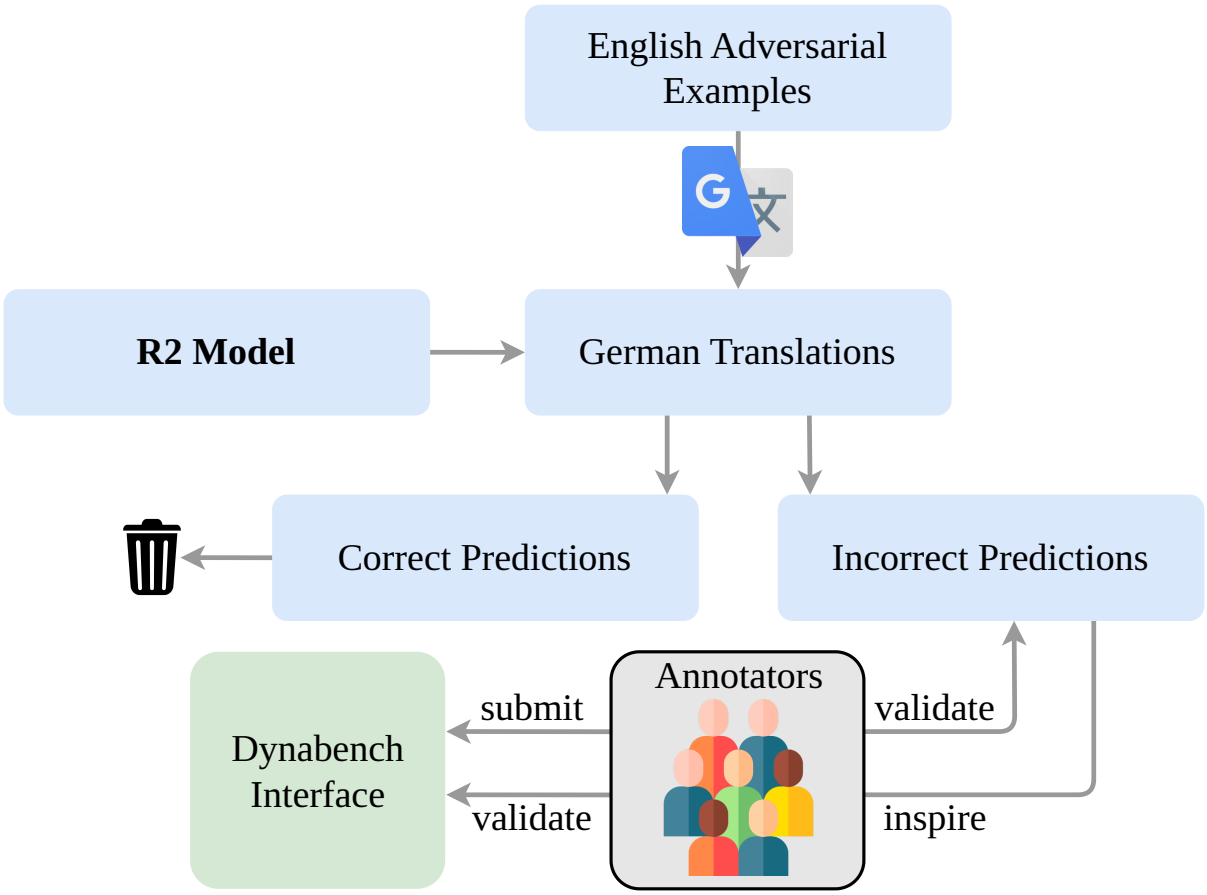

Figure 5: DADC workflow for R2, where we let annotators validate model tricking translations of English adversarial examples.

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 9 and 10.

5.2 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.

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

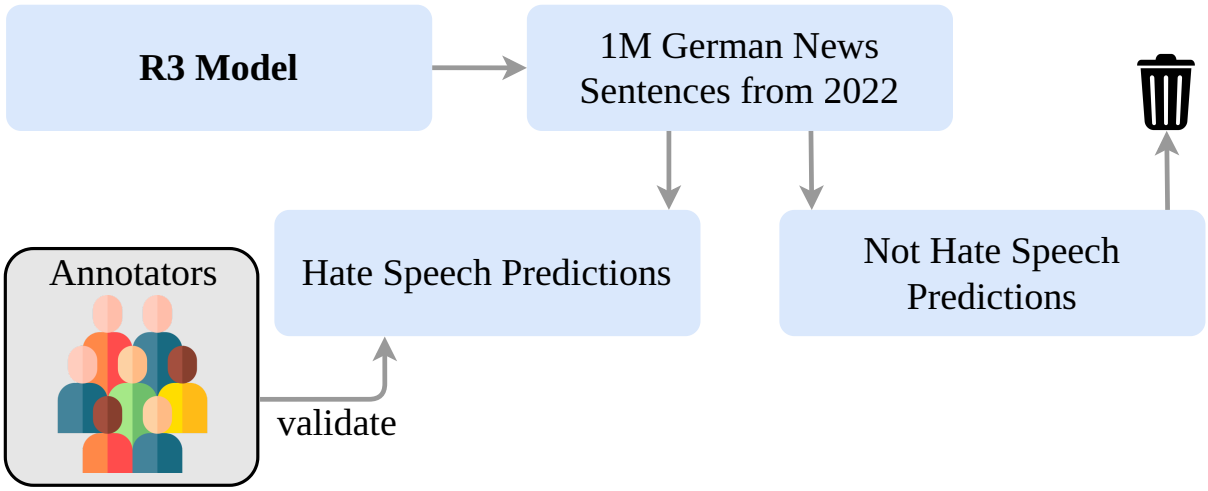

Figure 6: Workflow of R3, where we task annotators with validating model tricking newspaper sentences.

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's no indication of targeting any protected group, it doesn't 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.

5.3 Round 2: Translated Adversarial Examples

For R2, we translated English adversarial examples collected by Vidgen et al. (2021) to German using Google Translate[10] 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 who could accept or reject it. 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.

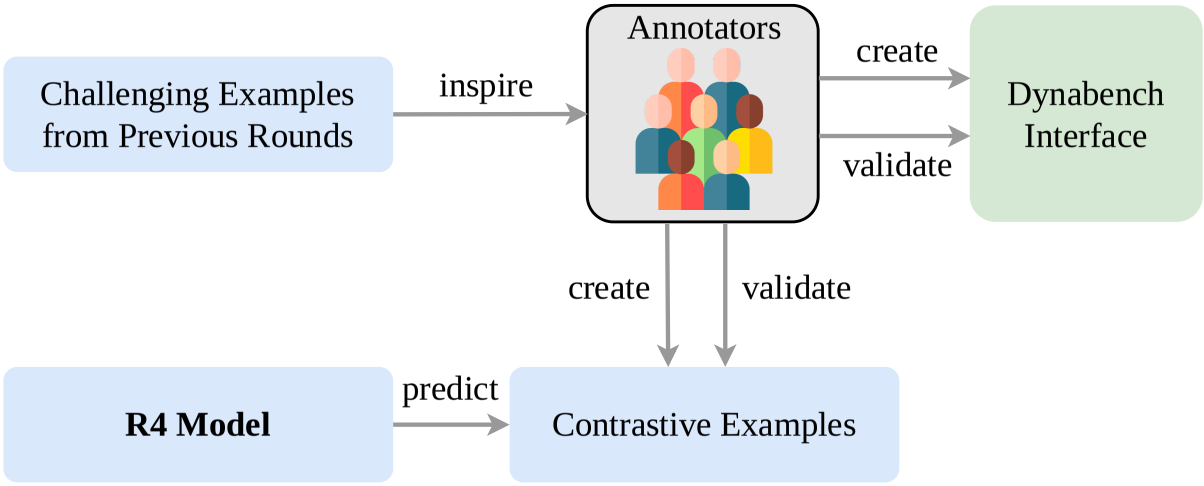

Figure 7: Workflow of R4, where we let annotators create contrastive examples to challenging entries from previous rounds.

5.4 Round 3: Newspaper Sentences

For R3, we used the sentences sampled from German newspaper articles published in 2022[11](Goldhahn et al., 2012). 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 1 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.

5.5 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 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 *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, faulty, 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

[10]`https://translate.google.com`

[11]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	1,175	2,175
R2	3,043	1,091	4,134
R3	48	3,179	3,227
R4	575	885	1,460
Total	4,666	6,330	10,996

Table 2: 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 3: Label distribution in GAHD across data splits.

dataset.

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.

5.6 Full Dataset

The final dataset contains 10,996 examples, with 4,666 (42.4%) labeled as hate speech. Table 2 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 3.

Model Error Rate In R1, annotators successfully tricked the target model with 41.3% of entries. 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 they were only validated and included in the dataset for having fooled the target model.

Inter-Annotator Agreement We observed some variation of inter-annotator agreement between the rounds and but overall relatively high agreements and provide two points of discussion: (1) 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. (2) 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 the dataset. 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,

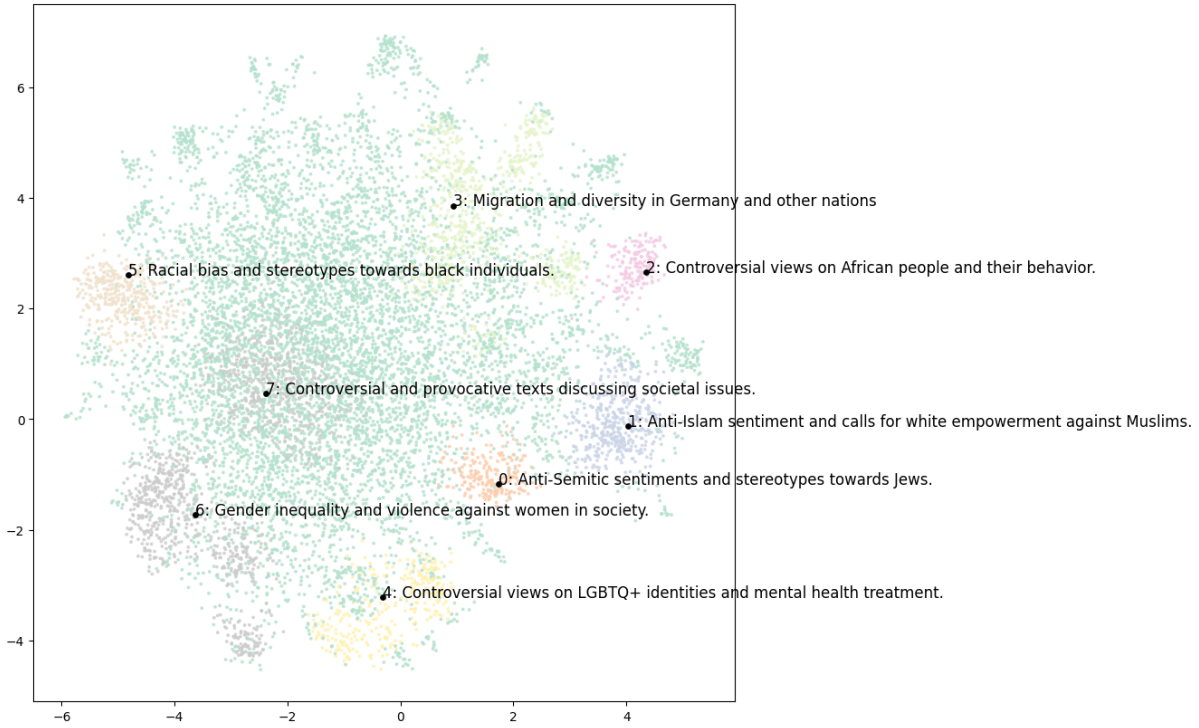

Figure 8: An overview of the most important topics in GAHD.

we use GPT3.5-turbo[12] to generate cluster descriptions based on the top words (ranked via TF-IDF) and sentences of the cluster.

We obtain eight clusters ranging from ca. 150 to over 700 examples, with over 7,000 comments remaining uncategorized. Figure 8 shows the clusters visualised in 2D. We observe that the clustering leads to a categorization into major protected groups and that it highlights in which discourse context the specific protected group is typically attacked. For example, the description of the cluster about LGBTQ+ people connects this topic to a mental health discourse, indicating that entries in the dataset might attack LGBTQ+ people by viewing their identities as "treatable mental health issues".

6 Experiments

6.1 Does the Dataset Improve Model Robustness?

We want to test to what degree the dataset improves robustness systematically. For that purpose, we train gelectra-large on the web-sourced datasets from Section 5.1, 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 5.1. 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 9 displays the results averaged over five random seeds. The shaded areas show the bootstrapped 95% confidence intervals around the average performance. Each new round clearly improves the performance on HateCheck with earlier rounds having a larger impact than later rounds. The performance on the GAHD test split improves as well, however, the last round,

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

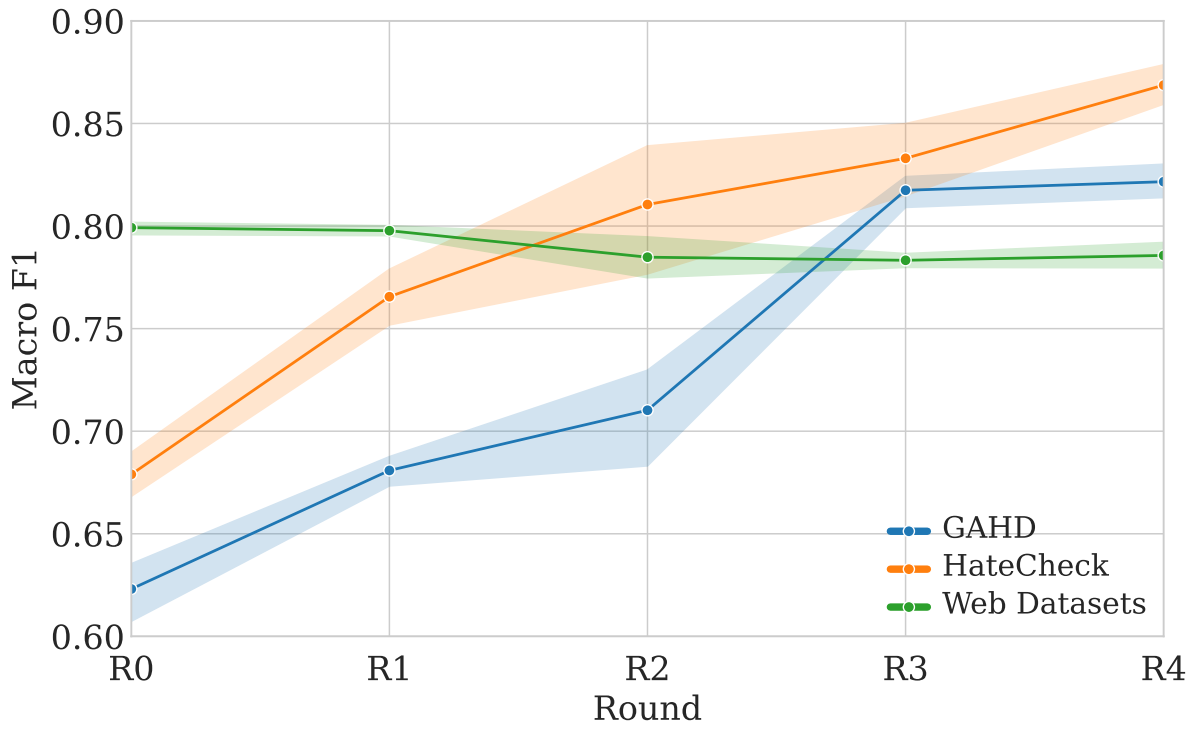

Figure 9: Model performance on different testsets as we add new training data across four rounds of DADC.

containing contrastive examples, has almost no impact. The macro F_1 on the initial datasets drops slightly, after including R2 data. Since the order of testing and size of each round affect the improvement per round, we control for those factors in the experiments in the next section.

6.2 Which Round Provided the Most Effective Examples?

To isolate the effect of each round and control for dataset size, we randomly sample 1,000 examples from the training split of each round and compare the effect of adding these to training splits of the web-sourced data. We use the same gelectra-large model and hyperparameters as in the previous section, and perform the experiments over five random seeds for sampling as well as model training.

Results Figure 10 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. The performance on GHAD and HateCheck varies between rounds, which contrasts the model the performance on the web datasets remaining mostly unchanged. Overall, we observe that mixing data from different rounds yields better results than only using data from a single round.

6.3 How Robust are Large Language Models and Commercial APIs?

To estimate how challenging GAHD is, 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.[13] (OpenAI, 2023) We also test the openly-available LeoLM models, which are based on Llama 2 Touvron et al.

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

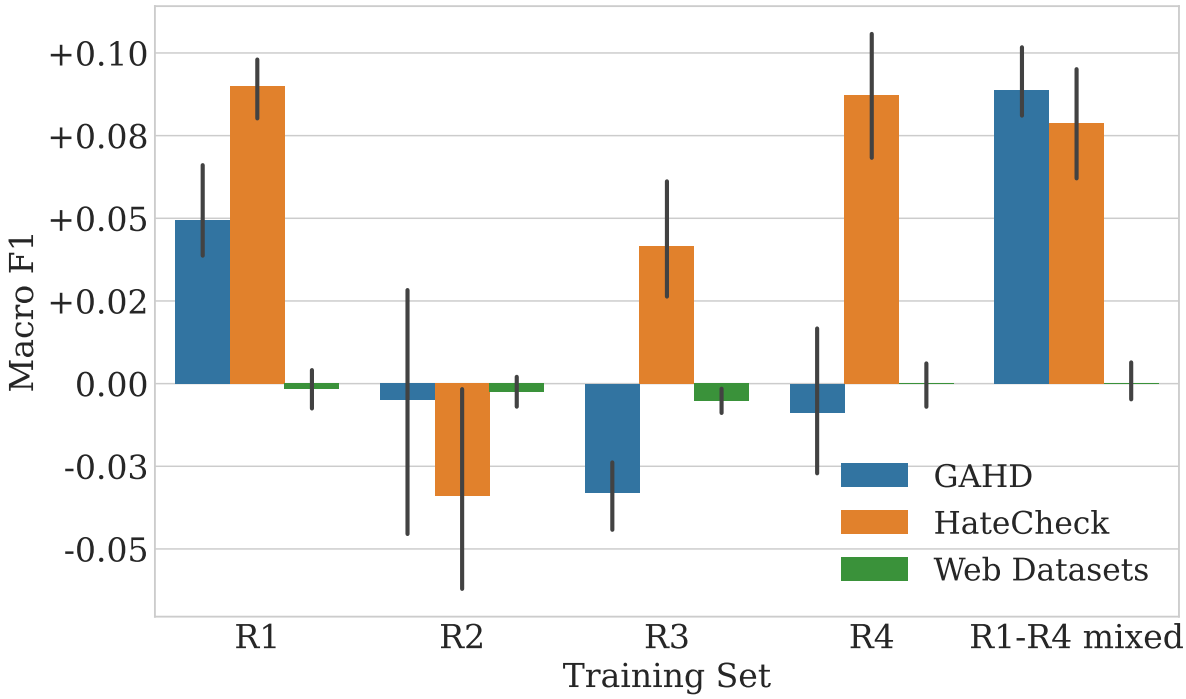

Figure 10: Impact on performance when including 1,000 adversarial examples in the training data.

(2023), and have been further pretrained and instruction tuned for German.[14] We evaluate all models in a zero-shot and five-shot scenario.

Content Moderation APIs The Perspective API by Google Jigsaw[15] and the content moderation API by OpenAI[16] 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 11 contains additional evaluation details.

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.

7 Conclusion

In this report, we gave an introduction to hate speech detection and then presented GAHD, a German hate speech detection dataset produced via four rounds of dynamic adversarial data collection. In rounds 2, 3, and 4 we explore new strategies for supporting the annotators by

[14] A paper about the LeoLM model suite has not yet been released. The training procedure is described in this blog post: `https://laion.ai/blog/leo-lm/`.

[15] `https://www.perspectiveapi.com/`

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

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
gelectra-large		
gelectra (web)		0.623
gelectra (web + GAHD)		0.822

Table 4: Macro F_1 of LLMs and content moderation APIs on the test set. We include the results of gelectra-large, our target model for comparison: *gelectra (web)* refers to gelectra-large fine-tuned only on web data, and *gelectra (web+GAHD)* refers to gelectra fine-tuned on web data and GAHD.

suggesting candidates for validation or inspiration. GAHD contains ca. 11,000 examples (42.4% hate speech), including 1,300 contrastive examples. Our experiments demonstrated that: (1) Training on adversarial data clearly improves robustness: strongly improved performance on GAHD and HateCheck with minimal or no loss on web-sourced data. (2) hand-crafted adversarial examples are more effective than collected and validated examples. (3) Training comparatively small models on in-domain adversarial data can make them more robust than large language models or commercial APIs.

Based on our findings, we highlight three key areas for future work:

Diversity in Adversarial Examples Mixing adversarial data from different rounds led to more consistently improved results than using data from only a single round. This indicates that diversity in how data is created or collected is important for a useful adversarial dataset. We thus believe that finding and testing more methods for supporting diverse adversarial data collection could further improve DADC.

Hand-Crafted vs. Validated Examples In our experiments, hand-crafted adversarial examples had more positive impact per example for increasing model robustness. Collected and validated adversarial examples had less positive impact per example, but they can be created more efficiently. Thus, the lower cost per example can offset or even reverse this disadvantage, as is demonstrated by the results in Figure 9. Future work could search for data collection methods that find a better trade-off between creation efficiency and per-example impact. Specifically, counterfactual data augmentation is a promising avenue for creating contrastive and adversarial candidates for annotators Sen et al. (2023).

Robustness vs. In-Domain Accuracy While training on adversarial data improved robustness, it did not increase performance on the web-sourced datasets. This result is not surprising, since social media-sourced datasets, including their test splits, contain exactly the sampling biases that we aim to avoid and counteract with adversarial datasets. Both, web-sourced datasets, and synthetic ones, like ours, do not represent the input distribution encountered after deployment. It remains an open question how to bridge this gap effectively.

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 the dataset, 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.

References

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

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. $\texttt{RP-Mod}\ \&\ \texttt{RP-Crowd:}$ moderator- and crowd-annotated german news comment datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=NfTU-wN8Uo.

Eric Barendt. What is the harm of hate speech? *Ethical Theory and Moral Practice*, 22(3): 539–553, 2019. ISSN 13862820, 15728447. URL http://www.jstor.org/stable/45217319.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1023.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 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, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.696. URL https://aclanthology.org/2021.emnlp-main.696.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 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, July

2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.275. URL https://aclanthology.org/2022.naacl-main.275.

Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 12 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00041. URL https://doi.org/10.1162/tacl_a_00041.

Rui Cao and Roy Ka-Wei Lee. 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), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.557. URL https://aclanthology.org/2020.coling-main.557.

Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.598. URL https://aclanthology.org/2020.coling-main.598.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 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), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.14. URL https://aclanthology.org/2022.woah-1.14.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1461. URL https://aclanthology.org/D19-1461.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

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

John Rupert Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Philological Society, Oxford, 1957.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), jul 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Jun. 2018. doi: 10.1609/icwsm.v12i1.14991. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14991.

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. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 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, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Edita Grolman, Hodaya Binyamini, Asaf Shabtai, Yuval Elovici, Ikuya Morikawa, and Toshiya Shimizu. 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, 2022. Association for Computing Machinery. ISBN 9781450392075. doi: 10.1145/3503252.3531309. URL https://doi.org/10.1145/3503252.3531309.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 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, 2018. Association for Computing Machinery. ISBN 9781450360043. doi: 10.1145/3270101.3270103. URL https://doi.org/10.1145/3270101.3270103.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234.

Mika Hietanen and Johan Eddebo. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458, 2023. doi: 10.1177/01968599221124309. URL https://doi.org/10.1177/01968599221124309.

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

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 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), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.17. URL https://aclanthology.org/2022.woah-1.17.

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. 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, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 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, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.woah-1.12.

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

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 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, 2019. Association for Computing Machinery. ISBN 9781450377508. doi: 10.1145/3368567.3368584. URL https://doi.org/10.1145/3368567.3368584.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 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, 2021. Association for Computing Machinery. ISBN 9781450389785. doi: 10.1145/3441501.3441517. URL https://doi.org/10.1145/3441501.3441517.

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

Pasquale Minervini and Sebastian Riedel. 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, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1007. URL https://aclanthology.org/K18-1007.

Melody Moh, Teng-Sheng Moh, and Brian Khieu. 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, 2020. doi: 10.1109/IMCOM48794.2020.9001767.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, January 2022. ISSN 2198-6053. doi: 10.1007/s40747-021-00608-2. URL https://doi.org/10.1007/s40747-021-00608-2.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Rajvardhan Oak. 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, 2019. Association for Computing Machinery. ISBN 9781450367479. doi: 10.1145/3319535.3363271. URL https://doi.org/10.1145/3319535.3363271.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 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, July

2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.173.

OpenAI. GPT-4 technical report, 2023.

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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Agnieszka Pluta, Joanna Mazurek, Jakub Wojciechowski, Tomasz Wolak, Wiktor Soral, and Michal Bilewicz. Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, 13(1):art. no. 4127, 2023.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523, jun 2021. ISSN 1574-0218. doi: 10.1007/s10579-020-09502-8. URL https://doi.org/10.1007/s10579-020-09502-8.

Nils Reimers and Iryna Gurevych. 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, 11 2019. URL https://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. 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, 11 2020. URL https://arxiv.org/abs/2004.09813.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 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), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.15. URL https://aclanthology.org/2022.woah-1.15.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13.

Magnus Sahlgren. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm University, 2006.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "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, May 2021. doi: 10.1609/icwsm.v15i1.18085. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18085.

Rupak Sarkar and Ashiqur R. KhudaBukhsh. Are chess discussions racist? an adversarial hate speech data set (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):15881–15882, May 2021. doi: 10.1609/aaai.v35i18.17937. URL https://ojs.aaai.org/index.php/AAAI/article/view/17937.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.347. URL https://aclanthology.org/2022.naacl-main.347.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.649.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

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

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3509. URL https://aclanthology.org/W19-3509.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

Jeremy Waldron. *The harm in hate speech.* Harvard University Press, Cambridge, Mass, 2012.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics:*

ACL 2022, pages 202–217, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.18. URL https://aclanthology.org/2022.findings-acl.18.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL https://aclanthology.org/N19-1060.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117, 07 2019. ISSN 0007-0955. doi: 10.1093/bjc/azz049. URL https://doi.org/10.1093/bjc/azz049.

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. 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, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

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

Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, jun 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.598. URL https://peerj.com/articles/cs-598. Publisher: PeerJ Inc.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

8 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.

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.

Potential Misuse 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 haven't 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.

9 Initial Datasets

able 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.

10 Target Model Training Details

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

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

Table 6: Hyperparameters of the target model.

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 use 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 one target model on one GPU for five epochs took approximately 30 minutes, we estimate that our experiments overall ran for ca. 30 GPU hours. We used GitHub Co-Pilot and ChatGPT for coding assistance.

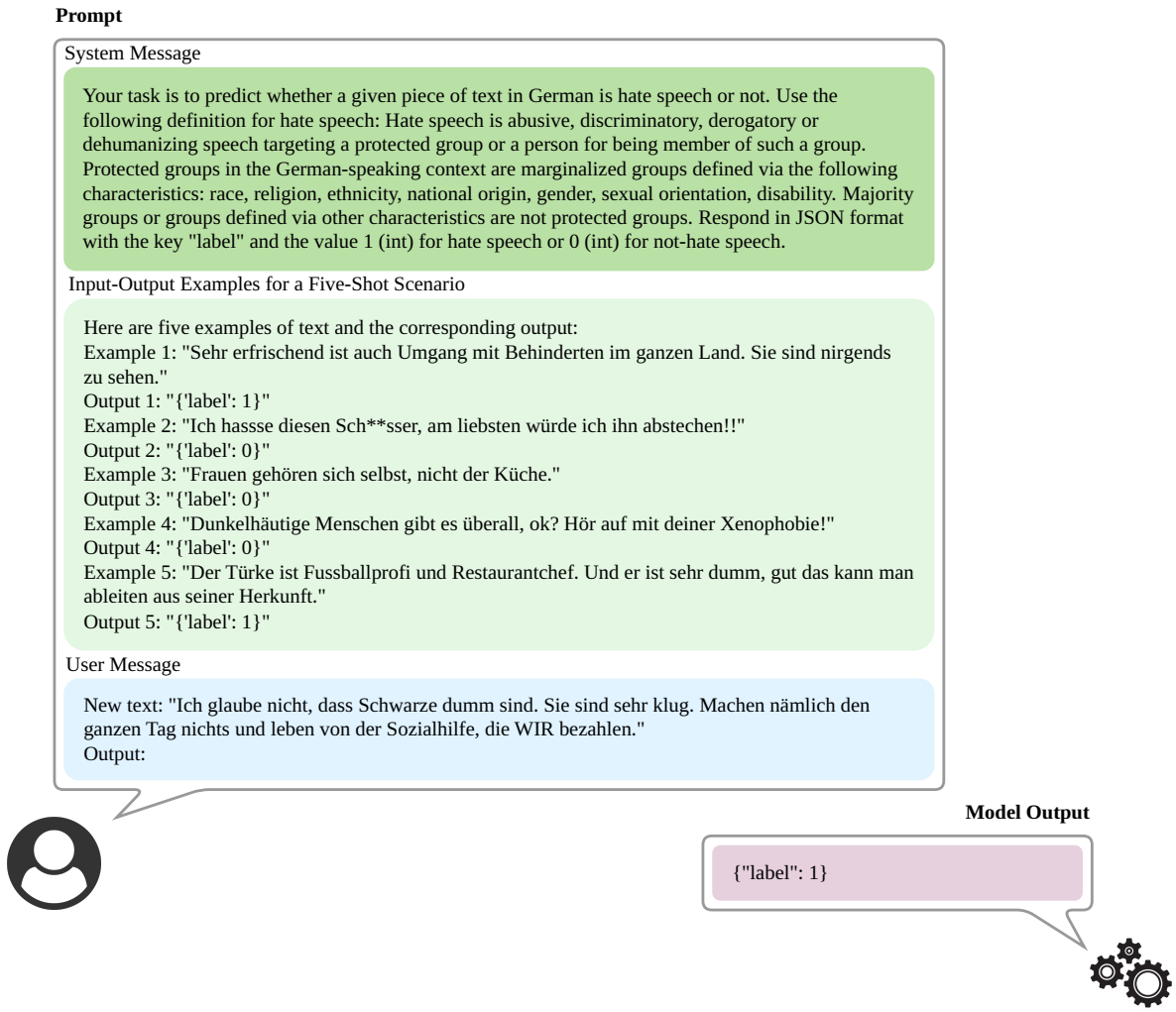

Figure 11: Five-shot prompt for GPT models. The prompt contains real examples from GAHD. For LeoLM models the last sentence of the system message is changed to "Respond in with 'True' (for hate speech) or 'False' (for not-hate speech).". We change the example outputs in the five-shot scenario accordingly scenario are changed accordingly.

11 Evaluation of Large Language Models and APIs

Here, we provide additional details for the evaluation settings in Section 6.3:

Large Language Models We evaluated all LLMs with the same prompt containing a task description, a hate speech definition, and a response format. Figure 11 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[17] which guarantees that the models generate valid JSON. However, the LeoLM models were not able to respond consistently valid JSON. We thus changed the response format for LeoLM to only only one token: *TRUE* or *FALSE*. We set the generation length to 1 ensured that both tokens are present in the LeoLM vocabulary. If a LeoLM model responded with a different token we regenerated the response.

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

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[18] 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.

12 Data Statement

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

12.1 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 permissevely licensed.

12.2 LANGUAGE VARIETY

We instructed the annotators to create texts in standard German. Newspapers in German-speaking countries often require comment section 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.

12.3 SPEAKER DEMOGRAPHICS

The dataset 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 is hard to characterize, as it contains sentenes from a wide range of news websites. From that fact, we can assume that the speaker demographics mostly consists of German journalists. However, as described in Section 5.4, we found some sentences that rather look like newspaper comments sentences out of a newspaper article.

12.4 ANNOTATOR DEMOGRAPHICS

Section 4.4 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. Four were female (57%), and three were male (43%). Three annotators had a high school diploma and were currently pursuing a bachelor's degree (43%), three a had a bachelor's degree 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 that worked at the university. Both were male, had a master's degree, were native German speakers, and in

[18] See: `https://perspectiveapi.com/`

the age ranges of 30 to 39, and 40 to 49. The lead author took the role of expert annotator. We redacted the demographics of the expert annotator to remain anonymous for the peer review.

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 course work or student projects.

We paid the annotators well above the local minimum wage, and according to university guidelines. The DADC rounds were spread 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 annoators then worked remotely. We gave the annotators further feedback and instructions via online meetings.

12.5 SPEECH SITUATION

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

12.6 TEXT CHARACTERISTICS

We describe the label distribution and general topics present in GAHD in Section 5.6.