Oddballs and Misfits: Detecting Implicit Abuse in Which Identity Groups are Depicted as Deviating from the Norm

Michael Wiegand[*†]
[*]Digital Philology
Faculty of Philological and Cultural Studies
University of Vienna
AT-1010 Vienna, Austria
michael.wiegand@univie.ac.at

[†]Digital Age Research Center
University of Klagenfurt
AT-9020 Klagenfurt, Austria
michael.wiegand@aau.at

Josef Ruppenhofer
Center of Advanced Technology for Assisted Learning and Predictive Analytics
FernUniversität in Hagen
D-58097 Hagen, Germany
josef.ruppenhofer@fernuni-hagen.de

Abstract

Warning: *This paper contains content that may be offensive or upsetting.*

We address the task of detecting abusive sentences in which identity groups are depicted as deviating from the norm (e.g. *Gays sprinkle flour over their gardens for good luck*). These abusive utterances need not be stereotypes or negative in sentiment. For this type of abuse, we are the first to present a study on how to detect it. We introduce datasets for this task created via crowdsourcing that include 7 different identity groups. We also report on classification experiments and show that only large language models detect this abuse reliably.

1 Introduction

Abusive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Examples are (1)-(2).

(1) stop editing this, you dumbass.
(2) you stupid fucking idiot, fucking kill yourself

Closely related terms, e.g. *cyber bullying* (Zhong et al., 2016) or *hate speech* (Waseem and Hovy, 2016), are compatible with the definition above.

Due to the rise of abusive language online, a common application is **content moderation** on social media, where NLP methods guide human reviews to the most relevant microposts.

The detection of **implicitly abusive language** (Waseem et al., 2017), i.e. abusive language **not conveyed by unambiguously abusive words** (e.g. *scum*, *tosser*), remains challenging (van Aken et al., 2018; Wiegand et al., 2021b; Ocampo et al., 2023):

(3) Did Stevie Wonder choose these models?
(4) You look like the back end of a bus.

Implicitly abusive language often targets **identity groups**, i.e. groups of people united by a common characteristic, fostering a sense of community (e.g. *Jews*, *gay people*, *women* etc.). Abusive language towards such groups can often be regarded as a form of **othering**, i.e. a means of stigmatizing the target as not fitting in within the norms of a social group (Burnap and Williams, 2016).

A common form of stigmatizing identity groups is by imposing negative stereotypes on them:

(5) Hispanics are **lazy** people.
(6) Muslims are **terrorists**.

Recent datasets on implicitly abusive language cover instantiations of such stereotypes fairly well (Sap et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021a,b; Hartvigsen et al., 2022; Wiegand et al., 2022). In this paper, we focus on the detection of a subset of abusive instances of othering that **do not coincide with negative sentiment and may not even necessarily represent stereotypes**:

(7) Gays sprinkle flour over their gardens for good luck.
(8) Some Muslims go to the barber for armpit haircuts.
(9) Asians do not use banks.
(10) Women do not get horny.
(11) Black people occasionally wear pyjamas in public.

(7)-(11) depict the identity groups as deviating from general norms of modern-day society. Such **norm-contraventions are frequently perceived as abusive by the targeted identity groups**. A classifier specializing in detecting this form of abuse could supplement existing classifiers for abusive language detection since, as we will show, the latter are unable to detect this type of abuse reliably.

[1] http://thelawdictionary.org

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2200–2218
November 12-16, 2024

In order to quantify the phenomenon, we manually annotated the abuse directed at identity groups on the dataset for implicit abuse by Ocampo et al. (2023), which is a union of 7 previous datasets. 65% of the (declarative) sentences can be considered instances of othering. While only 5% of all abusive instances lack negative sentiment, in 80% of these cases the identity groups are depicted as deviating from the norm. Thus, deviating from the norm represents a **prominent subset of the difficult non-negative abusive sentences**. Compared to the figures reported by Wiegand et al. (2021b) on other subtypes of implicit abuse, this phenomenon is similarly frequent as dehumanization, euphemisms or comparisons, all of which have previously been examined (Mendelsohn et al., 2020; Wiegand et al., 2021a, 2023).

We also had crowdworkers, all native English speakers without specific backgrounds, compare the severity of different examples of implicit abuse (20 examples for each type). The examples were presented in pairs without revealing their types. Crowdworkers had to decide which example they considered more severe. Overall, our novel type of abuse was judged even more severe than euphemisms or comparisons.[2]

Since our above sample is too small for a proper study we created 2 new English datasets: The first represents true-to-life examples extracted from Twitter. The second comprises sentences constructed by crowdworkers. Due to ethical concerns, the crowdworkers are **not** asked to form sentences targeting a specific identity group but an unspecific group of people represented by a 3rd person pronoun (e.g. *they*). We then instantiate these pronouns with identity groups and have the resulting sentences validated as abusive language by other crowdworkers. Thus, we establish that depicting a group of people as deviating from the norm is a **general property of implicitly abusive language** that can be observed across different identity groups.

We focus on sentences that can be interpreted **without any additional context**. The task is a **binary (sentence-level) classification problem** in which norm-compliant behaviour is to be distinguished from norm-contravening behaviour. We also report on classification experiments and show that this task benefits from recent language models.

Our **contributions** are the following:

- We introduce novel datasets for detecting behaviour or properties deviating from the norm.
- We show that identity groups being depicted in such way is perceived as abusive language.
- We demonstrate that such abuse cannot be detected effectively by previous classifiers.
- We propose a supervised classifier trained on text instances augmented by GPT-4.

All data created as part of this research are **available upon request**.

2 Related Work

Previous work on abusive language mostly follows a one-size-fits-all approach (Nobata et al., 2016; Badjatiya et al., 2017; Fortuna and Nunes, 2018). Surveys on existing datasets do not address implicit abuse (Vidgen and Derczynski, 2020; Poletto et al., 2021). However, the recent roadmap on implicit abuse by Wiegand et al. (2021b) identified as subtypes: dehumanization (Mendelsohn et al., 2020), call for action, multimodal abuse (Kiela et al., 2020), comparisons (Wiegand et al., 2021a), euphemistic abuse (Wiegand et al., 2023) and abuse towards identity groups (Hartvigsen et al., 2022). Our work aligns with the last subtype. Further subtypes are jibes (Sodhi et al., 2021), sarcasm and white grievance (ElSherief et al., 2021).

There has already been previous research related to othering beyond stereotypes: Burnap and Williams (2016) and Alorainy et al. (2019) analyze the *juxtaposition of 1st and 3rd person pronouns* to contrast the norm (*us*) with identity groups (*them*). Wiegand et al. (2022) examine *non-conformist views*, sentences expressing negative sentiment towards targeted groups. Our focus on othering differs as it is not limited to specific lexical items, like pronouns, nor solely to negative sentiment.

Our task is also related to *framing* (Mendelsohn et al., 2021; Ali and Hassan, 2022) since the presentation of identity groups selects aspects of a perceived reality and makes them more salient in a communicating text (Entman, 1993). These aspects do not have to apply to the identity groups in reality. Thus, they can also be considered *misinformation* (Zhou and Zafarani, 2020; Guo et al., 2022).

Our work is also anchored in social psychology: Lindström et al. (2017) observed that what is *common* (=norm-compliance) is often regarded as *moral* and that rare positive behaviour, e.g. altruism, is judged less moral than common positive behaviour. Our work echoes this sentiment,

[2]Details on this experiment are available in Appendix B.2.

suggesting that norm-compliance is perceived positively, while deviation from norms is viewed negative. Leary (2000) and Tangney and Dearing (2002) find that guilt and shame are social emotions typically experienced when individuals transgress a (social) norm. The audience of sentences depicting norm-contravention may likewise associate similar negative emotions with these utterances, potentially leading to strong disapproval. People are often motivated to punish those who violate societal norms, even if they are not personally affected by these violations (Fehr and Fischbacher, 2004; Buckholtz and Marois, 2012). Authors who engage in abusive language targeting norm violations may aim to trigger this punitive reflex in their audience. After all, *call for action* is a typical characteristic of implicitly abusive language. Finally, the fact that identity groups are the target of abusive sentences displaying norm-contravention might be explained by the fact that these groups are also sanctioned more frequently for norm violations than others (Wolbring et al., 2013; Winter and Zhang, 2018).

3 Data

Our new data represents a subtype of abuse that depicts identity groups as deviating from the norm. This is a form of othering. The utterances may be stereotypes but they do not have to be. They may also coincide with examples of framing or misinformation. We do not consider abusive utterances that are explicitly negative in sentiment. Such sentiment, which is conveyed by words unambiguously negative in meaning,[3] e.g. *poor* or *sad*, has been dealt with in previous work (Wiegand et al., 2022).

We create **2 datasets**: Due to the rareness of this phenomenon, we created a **constructed dataset**, i.e. a dataset in which sentences are invented (§3.1). In this way, we can cover various areas of life. However, since that dataset does not necessarily reflect texts in social media, we also produce a **Twitter dataset** comprising attested sentences (§3.2).[4]

The annotation of all datasets was produced via crowdsourcing. As a platform we used Prolific academic.[5] We did not specifically target any particular profession, age group, gender or ethnicity for our sample. We randomly sampled crowdworkers from the pool of available crowdworkers on Prolific who reside in Western countries (i.e. USA, UK, Ireland, Australia and Canada) and made sure that English is their first language. We focused on crowdworkers in Western countries because our work exclusively examines behaviors that are compliant with or contrary to Western norms. All crowdworkers had to have an overall approval rate of 100%. Our annotation tasks were divided into smaller segments for the crowdworkers, such as creating 30 sentences or evaluating the labels of 100 sentences. This approach led to the participation of a diverse group of over 100 crowdworkers. Considering the large number of participants, the platform's extensive base of several thousand native English speakers in Western countries, and the absence of further demographic restrictions, we are confident that our ratings offer a representative cross-section of Western society.

3.1 Constructed Dataset

Following previous work on creating a dataset for a rare subtype of implicit abuse (Wiegand et al., 2021a), we asked crowdworkers to *invent* instances of our targeted phenomenon, i.e. behaviour or properties that deviate from the norm. However, due to ethical reasons we did not ask them to think of *identity groups* but an unspecific group of people that should be referred to by a 3rd person plural pronoun. Therefore, the resulting sentences do not attack any identity group. Subsequently, we instantiated the pronoun with specific identity groups.

Our dataset is created through a series of individual tasks. In each task, a single crowdworker had to invent about 30 sentences or judge the label of about 100 sentences. Figure 1 illustrates the order of those tasks which we also describe in the following. We repeatedly ran through this pipeline until no more new sentences were obtained.

❶ **Norm-Contravening Sentences.** Crowdworkers were asked to invent sentences in which a generic group of people (represented by a 3rd person plural pronoun) was depicted as displaying a behaviour or property that deviates from the norm. **By norm we understand behaviors or situations deemed typical within modern-day society**, with a particular focus on Western societies, as this cultural backdrop is predominant in the English language data we are using. Our attention is on **social norms** (*Wear black to a funeral*) and **conventions** (*Follow the rules of English grammar*) rather than moral or legal norms (Elster, 2007), e.g. honesty or

[3] More information on sentiment is given in Appendix B.3.

[4] We consciously avoided sampling sentences from SOCIAL-CHEM-101 (Forbes et al., 2020) as that work is strongly related to Moral Foundations (Haidt, 2012) and the predominant norms in the dataset demonstrate explicit sentiment.

[5] www.prolific.com

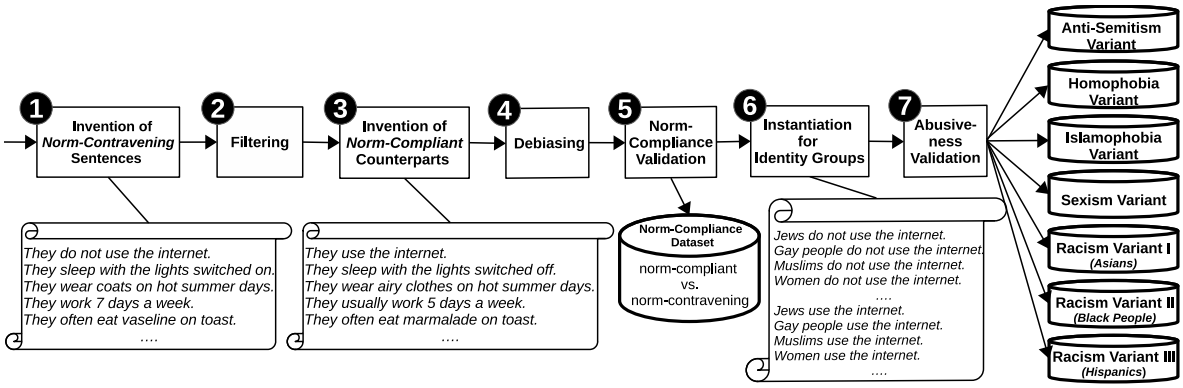

Figure 1: Illustration of how the **constructed dataset** (i.e. norm-compliance dataset and its 7 variants) is created.

justice. In other words, we are interested in norms whose violation might at most result in shame but not in blame (Malle et al., 2014) since in our exploratory experiments we observed that the latter often coincides with explicit sentiment, which we avoid in this work as stated above.

❷ **Filtering.** The sentences produced by the crowdworkers required manual filtering by one co-author. This involved removing near duplicates and occasional cases of explicitly negative sentiment.

❸ **Norm-Compliant Counterparts.** For each *norm-contravening* sentence, one co-author manually created a sentence in which the depicted behaviour or property follows the norm. We refer to this as *norm-compliant* sentences. These instances were created as *contrast sets* (Gardner et al., 2020; Li et al., 2020; Sen et al., 2022), i.e. sentences that are structurally similar to the *norm-contravening* sentences. Thus, we obtain a difficult dataset in which *norm-compliant* and *norm-contravening* sentences are hard to distinguish from each other. By having a co-author rather than crowdworkers create those sentences, we follow Gardner et al. (2020) who recommend such data to be created by experts.

❹ **Debiasing.** By computing the Pointwise Mutual Information between words and the two classes of our norm-compliance dataset, i.e. *norm-compliant* and *norm-contravening*, we identified a set of spurious correlations (Ramponi and Tonelli, 2022). Most of them were caused by the way we created *norm-compliant* sentences: In order to change a *norm-contravening* sentence (12) minimally to a *norm-compliant* one, often simply some adverbial was removed or added (13). As a result, words, such as *rarely*, were biased towards the class *norm-compliant*. However, these words should not be predictive for this class, as *norm-contravening* sentences, such as (14), are equally possible. Therefore, we replaced these sentences by other sentences not containing these words (15).[6]

(12) They wash clothing by hand. (*norm-contravening*)
(13) They **rarely** wash clothing by hand. (*norm-compliant*)
(14) They **rarely** wear shoes outside. (*norm-contravening*)
(15) They wash clothing in washing machines. (*norm-compliant*)

❺ **Norm-Compliance Validation.** 5 different crowdworkers were asked to validate whether a given sentence represents a behaviour or property that deviates from the Western norm. Crowdworkers could also label a sentence as not being proper English. Only sentences in which the majority agreed on a label were used for further processing.

❻ **Instantiation.** We produced 7 variants of each sentence in which we replace the 3rd person pronouns by identity groups that represent frequent targets of abusive language, i.e. *Asians*, *Black people*, *gay people*, *Hispanics*, *Jews*, *Muslims* and *women*.

❼ **Abusiveness Validation.** Another set of crowdworkers were to rate the given instantiated sentences, both norm-compliant and norm-contravening, as *anti-Semitic*, *homophobic*, *Islamophobic*, *racist*, *sexist* or *not abusive*. **Only crowdworkers belonging to the identity group mentioned in a given sentence were to rate that sentence.** Often, the affected identity groups are the most competent to detect this type of abusive language (Pei and Jurgens, 2023). The crowdworkers could also flag a sentence as *improbable* if they considered it unlikely to be found on the Web. A sentence was excluded as soon as one crowdworker flagged it as improbable. The final label corresponds to the majority of the 5 crowdworkers.

[6] Appendix C provides some more details.

norm-compliance dataset		**variant 1: anti-Semitism**	
target	*they*	target	*Jews*
total sentences	1705	total sentences	1307
norm-contravening	901	anti-Semitic	621
norm-compliant	804	not abusive	686
		correspondence	90.9%
variant 2: homophobia		**variant 3: Islamophobia**	
target	*gay people*	target	*Muslims*
total sentences	1228	total sentences	1093
homophobic	500	Islamophobic	377
not abusive	728	not abusive	716
correspondence	88.2%	correspondence	86.6%
variant 4: sexism		**variant 5: racism (I)**	
target	*women*	target	*Asians*
total sentences	992	total sentences	1206
sexist	359	racist	419
not abusive	633	not abusive	787
correspondence	82.3%	correspondence	82.9%
variant 6: racism (II)		**variant 7: racism (III)**	
target	*Black people*	target	*Hispanics*
total sentences	1114	total sentences	1131
racist	396	racist	348
not abusive	718	not abusive	783
correspondence	84.7%	correspondence	84.3%

Table 1: Statistics on the **constructed dataset**.

The Final Dataset. Table 1 provides a statistic on the norm-compliance dataset and its variants. The size of the variants is smaller than the norm-compliance dataset as many instantiated sentences were judged improbable and thus removed. Table 1 also lists for each variant the **correspondence** of class labels to the norm-compliance dataset, i.e. we ascertain the proportion of sentences labeled as abusive (e.g. *racist*, *sexist*) originally labeled as *norm-contravening*, and sentences labeled as *not abusive* originally labeled as *norm-compliant*. On average, these labels correspond in 85.7% of the sentences. This indicates that the clear majority of sentences in which an identity group is depicted as deviating from the norm is judged abusive.

Table 2 lists the proportion of the areas of life (established via manual annotation) that are covered in our norm-compliance dataset illustrating the diversity of the dataset.

A random sample of 200 sentences of each part of our dataset was also annotated by one co-author. We compared these labels with the crowdworkers' majority vote. Though the co-author does not belong to any of the 7 identity groups, we still got a substantial agreement (Landis and Koch, 1977).[7]

3.2 Twitter Dataset

For our second dataset, we sampled sentences from Twitter in which one of the 7 identity groups from

[7]Appendix B.4 lists all individual agreement scores.

area of life	**%**	**area of life**	**%**	**area of life**	**%**
habits	17.1	social interaction	6.5	anatomy	2.9
food & drink	12.2	hygiene	5.0	mobility	2.9
views	11.2	skills	4.1	living	2.8
(dis)likes	10.4	family	4.0	technology	2.5
clothing	9.0	job & education	3.8	possessions	2.2

Table 2: Frequent areas of life in the **constructed** (norm-compliance) **dataset** established via manual annotation.

general information		**distribution of targets**	
total sentences	1028	sentences on *Jews*	87
abusive*	555	sentences on *gay people*	211
not abusive	473	sentences on *Muslims*	168
correspondence	75.9%	sentences on *women*	151
		sentences on *Asians*	111
		sentences on *Black people*	184
		sentences on *Hispanics*	116

*: anti-Semitic, homophobic, Islamophobic, sexist or racist sentences

Table 3: Statistics on the **Twitter dataset**.

§3.1 is mentioned. We searched on the Twitter history rather than fetching tweets that are currently streamed. This was done since, for several of our identity groups, we were only able to find a very small number of distinct tweets (i.e. only a few) that met the restrictions we formulated (as outlined below) within a reasonable time frame (e.g. a few weeks).

In order to be in line with the sentences representing implicitly abusive language from our constructed dataset (§3.1), the sentences from Twitter were not to contain any explicit abuse (e.g. slurs) or explicit sentiment.

Following the observation by Wiegand et al. (2022) that the overwhelming number of abusive remarks on identity groups realize the identity group as the agent (i.e. logical subject) of some predicate (e.g. full verb), we used queries that extracted such sentences. This is typically achieved by using a pattern `identity_group adverb` as in *Jews typically*, *Jews only*, *Jews rarely* etc. Depending on the particular query, we obtained up to several hundred unique tweets which were subsequently annotated via crowdsourcing with respect to norm-compliance and abusiveness (Table 3).

We also focused only on sentences that can be understood out of context. Typical situations in which this is not the case are:

- The tweet has an image, video or sound file attached whose content needs to be considered in order to understand the tweet.
- The tweet is part of a larger thread; the content of the entire thread needs to be considered in order to understand the tweet.

- The tweet can only be understood by knowing some background information on the author (e.g. specific demographic information).

Our Twitter dataset only comprises sentences rather than complete tweets. To protect people's privacy, mentions of both usernames and real names were removed from the dataset. This removal process was conducted manually to ensure comprehensive detection and exclusion of such mentions. We did not substitute those mentions; rather, we entirely removed them.

All the above restrictions reduced the size of our sample by about 70%. Each sentence was rated by 5 crowdworkers that belong to the identity group mentioned. The label inventory corresponds to that of the constructed dataset. The final labels correspond to the majority vote. Table 3 provides some statistics on the resulting dataset. Its small size (1000 instances) can be explained by the fact that we consider a rare phenomenon and by Twitter's intensive efforts to remove hate speech.

This dataset is primarily created to study the detection of implicit abuse. It is not used to study the categorization of norm-compliance per se.

On a random sample of 200 sentences, we also measured a substantial agreement of $\kappa = 0.65$ between one co-author and the majority vote.

4 Set-Up for Transformers

Two transformers are used as learning methods: BERT (`bert-base-uncased`) (Devlin et al., 2019) and DeBERTa (`deberta-large`) (Hea et al., 2021). We compare BERT, a foundational model, with DeBERTa, which introduces advanced architecture and benefits from more extensive training data. We fine-tune the pretrained models on the given training data using the FLAIR-framework (Akbik et al., 2019) with the hyperparameter settings from Wiegand et al. (2022), a study closely related to ours. We always report the average over 5 training runs (+ standard deviation). ***Appendix A contains details on the settings of all classifiers.***

We also use large language models, such as GPT-4 (OpenAI et al., 2024), for getting state-of-the-art text completions as outlined in the following.

5 Norm-Compliance

In this section, we **focus on our constructed norm-compliance dataset**. The task is to distinguish between sentences in which a generic group of people (i.e. *they*) is presented as deviating from the norm (i.e. *norm-contravening*) and sentences in which they fall within the norm (i.e. *norm-compliant*). This task has not been addressed before.

short prompt	Is this common?
long prompt	Is this common in our Western society?

Table 4: Prompts for GPT-4 based classifiers.

5.1 Classifiers not Trained on Our Dataset

Sentiment Analysis. Norm-contravening behaviours and properties may sometimes be perceived in a negative way. This suggests that the class *norm-contravening* may bear some relation towards negative sentiment. As stated in §3, we refrained from including explicitly negative sentiment in our dataset since such utterances are sufficiently represented in previous datasets. However, there are still utterances in our datasets that convey *implicitly negative sentiment* (Deng et al., 2013; Ding and Riloff, 2018; Zhou et al., 2021), e.g. (16) or (17). (Based on our manual inspection of the data we estimate 17% of the sentences to convey an implicitly negative sentiment.) As a sentiment classifier to detect all instances of negative sentiment in our dataset, we use *TweetEval* (Barbieri et al., 2020). Predictions of such sentiment are considered a proxy of *norm-contravening* sentences.

(16) They urinate in the sink.
(17) They did not finish high school.

GPT-4::zero-shot. We use a prompt asking a given sentence to be classified directly by GPT-4 as exemplified by (18) and (19). We interpret a completion beginning with *Yes* as a prediction for class *norm-compliant* and one beginning with *No* for class *norm-contravening*. We examine 2 prompts that vary in specificity (Table 4) since already minor variations in prompts are known to cause notably different completions (Zhang et al., 2021).

(18) ***prompt:*** [Usually, they use cutlery to eat.]$_{\textit{sentence to classify}}$ [Is this common in our Western society?]$_{\textit{long prompt}}$
completion: Yes, using cutlery to eat is common in Western society.

(19) ***prompt:*** [They eat cereal with water.]$_{\textit{sentence to classify}}$ [Is this common in our Western society?]$_{\textit{long prompt}}$
completion: No, it is not common in Western society to eat cereal with water.

LLaMA-2. Since GPT-4 is proprietary, we also consider LLaMA-2 (Touvron et al., 2023) for zero-shot classification as an *open-weight* alternative.

5.2 Within-Dataset Classifiers

As classifiers directly trained on our dataset, we do not only fine-tune **BERT** and **DeBERTa** (§4) but we also use **logistic regression** trained on a bag of words, i.e. a classifier that only draws knowledge from lexical items observed in the training data.

Knowledge Base. We assume that what is considered norm-compliant should also be found in large general-purpose knowledge bases, to some extent. Therefore, we also implemented a baseline using ConceptNet (Speer et al., 2017). For each sentence, we extract concepts from ConceptNet with CoCo-Ex (Becker et al., 2021) which are converted into a *ConceptNet vector ensemble* (Speer et al., 2017). We average over these embeddings and concatenate them with embeddings from DeBERTa, i.e. our best transformer, for each sentence. We train a feedforward neural network on our dataset where all sentences are represented by the above embeddings in FLAIR. The test data of our dataset are represented in the same fashion.

GPT-4::aug. In our last method, we augment each sentence from our dataset with the respective completion obtained from the zero-shot approach (§5.1). Therefore, the resulting dataset maintains the original amount of instances, however, each instance consists of the original sentence and the completion. We then fine-tune and test a transformer on these augmented instances. Learning on the text augmented by the GPT-4 completions may give a classifier additional helpful clues.

Given that this augmentation process results in instances possessing a greater textual length than the original, we also investigate whether the classifier's performance improvement is merely a function of longer text inputs. To address this, we generate a **control** configuration in which we augment each original sentence by some paraphrase so that the resulting text matches the length of the above augmented instances. Our prompt for creating a paraphrase (i.e. merely a hyphen) follows the specification from Wiegand et al. (2023).

5.3 Evaluation

Table 5 shows the performance on distinguishing between the classes *norm-compliant* and *norm-contravening*. We report macro-average precision, recall and F1-score for all experiments in this paper. For within-dataset classifiers (§5.2), we carried out a 5-fold cross-validation. As an upper bound, we tested a **human classifier** in which the judgment of one individual annotator was randomly sampled from the crowdsourced gold-standard annotation.

classifier	**Prec**	**Rec**	**F1** (*std*)
majority-class classifier	26.4	50.0	34.5
log. regr. trained on norm-compliance dataset	48.8	48.9	48.8
Sentiment Analysis (TweetEval)	51.9	51.4	51.7
Knowledge Base (ConceptNet)	64.6	64.4	64.5
LLaMA-2 (*short prompt*†)	73.3	58.6	65.1
LLaMA-2 (*long prompt*†)	74.7	60.4	66.8
BERT trained on norm-compliance dataset	68.7	68.7	68.7 (0.5)
DeBERTa trained on norm-compliance dataset	83.4	83.4	83.4 (0.4)
GPT-4::zero-shot (*short prompt*†)	84.3	83.1	83.7
DeBERTa trained on GPT-4::aug (*'control'*)	85.6	85.6	85.6 (0.8)
GPT-4::zero-shot (*long prompt*†)	86.7	85.0	85.8
DeBERTa trained on GPT-4::aug (*long pr.*†)	**93.3**	**93.3**	**93.3** (0.1)
human classifier	94.2	94.2	94.2

Table 5: Classification between *norm-compliant* and *norm-contravening* sentences on the constructed norm-compliance dataset (†: see Table 4).

Table 5 shows that our task requires models that incorporate world knowledge in addition to labeled training data. Logistic regression performs poorly. So does sentiment analysis. Plain language models perform notably better. Specifically, DeBERTa, the more sophisticated model, outperforms BERT.

The zero-shot classifiers using GPT-4 produce high scores. They outperform the classifiers based on LLaMA-2. A longer prompt (Table 4) is more effective than a shorter one probably since the latter lacks specificity as to the context of the norm.

The best performance is obtained by fine-tuning DeBERTa on training data augmented by GPT-4. The control configuration of the augmentation is significantly worse than the one using a long prompt. This suggests that improving performance due to text augmentation depends on adding predictive textual information. By manually inspecting the completions of the long prompt by GPT-4 (Table 4), we found that for utterances involving negation, which in our dataset represent 34% of the instances, the model does not choose one scope consistently. For example, in (20) the completion considers the wide scope that includes the negation of the sentence to classify, while in (21) it considers a narrow scope in which the negation is *not* included. GPT-4::zero-shot derives a wrong categorization from (21), since that classifier assumes the wide scope. Being trained on the concatenation of the original sentence and the completion, GPT-4::aug is able to learn what type of scope an individual completion replies to. Thus, GPT-4::aug can be much more accurate than GPT-4::zero-shot.

	log. regression	BERT	DeBERTa
automatic	80.4	81.7 (0.4)	88.2 (0.6)
without debiasing	67.3	77.1 (0.7)	87.0 (0.4)
proposed method	48.8	68.7 (0.5)	83.4 (0.4)

Table 6: F1 of standard classifiers on the constructed dataset with different *norm-compliant* sentences.

(20) ***sentence to classify:*** They do not have any bugs as pets.
completion: Yes, it is common in Western society to not have bugs as pets.

(21) ***sentence to classify:*** They do not wear rollerblades everywhere.
completion: No, it is not common in Western society to wear rollerblades everywhere.

5.4 Difference of *Norm-Compliant* Sentences

We now demonstrate that our **proposed method** to produce *norm-compliant* sentences by composing sentences that are structurally similar to the *norm-contravening* sentences and are additionally debiased (§3.1), results in a fairly difficult dataset.

We compare our proposed method against 2 simpler alternatives in which different *norm-compliant* sentences are employed (in both the training and the test data). In the first method, we obtain such sentences in an **automatic** way, namely as completions from GPT-4 by using the *norm-contravening* sentences as prompts and asking the model to produce a *norm-compliant* counterpart (22).

(22) ***prompt:*** [They eat rodents.]$_{\textit{norm-contravening sentence}}$ This is not common in our Western society. What would be common instead? They ...
completion: ... eat chickens, cows, pigs, or fish.

The second method simply takes our manually compiled *norm-compliant* sentences **without debiasing** (§3.1). Thus, this dataset still contains spurious correlations (as discussed in ❹ of §3.1).

Table 6 shows the performance of the two plain transformers and logistic regression on the 3 datasets that differ in the *norm-compliant* sentences. For all learning algorithms, the classification scores are notably higher for the two alternatives. While our proposed method represents a dataset in which, in terms of the surface realization, the sentences in the two classes hardly differ (Figure 1), this is not true for the alternatives in which there are biases that make automatic classification unrealistically simple: In the dataset containing the automatically generated *norm-compliant* sentences by GPT-4, the negative sentences are notably longer than the positive sentences (i.e. 11.3 vs. 7 tokens per sentence[8]).

[8]In our debiased dataset, sentences in both classes comprise the same average number of tokens.

In the biased version of the dataset, there are spurious correlations between words and the classes as already discussed in ❹ of §3.1.

6 Norm Deviation and Abusive Language

6.1 Experiments on the Constructed Dataset

We now evaluate classifiers to predict abusive language on the 7 variants of our constructed dataset.

Classifiers not Trained on Our Dataset. We consider 2 publicly available tools: **PerspectiveAPI**,[9] i.e. a tool for the general detection of abusive language, and the most recent transformer for implicitly abusive language detection focusing on identity groups from Hartvigsen et al. (2022), i.e. HateBERT fine-tuned on **ToxiGen**.

Moreover, we fine-tune DeBERTa on **ISHate** (Ocampo et al., 2023), a dataset consolidating 7 existing datasets for implicit abuse. Further, we fine-tune DeBERTa on the recent dataset for **euphemistic abuse** (Wiegand et al., 2023). We train the latter classifier on the abusive subtype *unusual properties* (23)-(24) since, though addressing individuals rather than identity groups, it is related to abusive language that depicts people as deviating from the norm. We want to examine the extent to which the two types of abusive language coincide.

(23) Your main hobby must be letting life pass you by.
(24) Your heart made an iceberg look warm.

We also re-use GPT-4::zero-shot from §5.1 where predictions of *norm-contravening* sentences are considered as predictions of abusive language.

Within-Dataset Classifiers. We employ the best classifier from §5, i.e. DeBERTa trained on GPT-4::aug, and also, for reference, plain DeBERTa from our previous experiments from §5. For these classifiers, we carry out a 5-fold cross-validation. However, we train on the instances of the norm-compliance dataset (Table 1), i.e. the dataset containing 3rd person mentions, and test on the respective instantiations of the 7 variants that focus on identity groups (predictions of *norm-contravening* sentences are considered abusive language). Thus, we can show that the knowledge to detect this abuse is not specific to a particular identity group.

Evaluation. Table 7 shows the classification results on the detection of abusive language. Classifiers not trained on our dataset mostly produce low

[9]https://perspectiveapi.com

classifier	anti-S.	homoph.	Islamo.	sexism	racism			average
					Asians	Black p.	Hispan.	
majority class	37.2	34.4	39.6	39.0	39.5	39.2	40.9	38.6
DeBERTa trained on euphemistic abuse	56.5 (2.0)	55.1 (1.3)	56.9 (2.4)	54.6 (1.7)	60.3 (1.3)	57.4 (1.7)	56.8 (0.9)	56.8 (1.6)
ToxiGen	60.5	59.6	61.5	60.7	58.2	54.7	57.3	58.3
DeBERTa trained on ISHate	58.3 (1.4)	56.9 (2.4)	61.3 (3.2)	59.4 (3.8)	60.1 (1.2)	60.3 (1.8)	62.9 (1.6)	59.9 (2.2)
PerspectiveAPI	57.0	59.5	59.8	58.7	64.6	75.6	65.4	62.9
DeBERTa trained on norm-compliance dataset	78.0 (2.3)	73.8 (1.3)	75.5 (0.7)	70.4 (3.3)	71.2 (2.7)	71.8 (1.2)	75.4 (0.8)	73.7 (1.8)
GPT-4::zero-shot (*long prompt*†)	82.1	76.1	75.5	78.2	72.8	75.2	73.9	76.2
DeBERTa trained on GPT-4::aug	**87.1** (0.3)	**79.2** (0.8)	**79.8** (0.4)	**78.7** (0.5)	**76.8** (0.6)	**77.1** (1.1)	**78.2** (0.7)	**79.6** (0.6)
human classifier	87.3	82.4	83.5	79.1	79.9	82.5	83.0	82.5

Table 7: F1 on abusive language detection on the constructed dataset (†: see Table 4).

classifier	F1 (*std*)
majority class	35.1
DeBERTa trained on euphemistic abuse	58.4 (2.2)
ToxiGen	60.8
DeBERTa trained on ISHate	61.5 (1.0)
PerspectiveAPI	68.3
DeBERTa trained on norm-compliance dataset	68.8 (0.8)
GPT-4::zero-shot (*long prompt*†)	73.0
DeBERTa trained on GPT-4::aug (*long prompt*†)	**75.1** (0.6)
human classifier	82.6

Table 8: Abusive language detection on the Twitter dataset (†: see Table 4).

classifier	%correct
LLaMA-2::zero-shot (*long prompt*†)	4.7
GPT-3.5::zero-shot (*long prompt*†)	51.8
GPT-4::zero-shot (*long prompt*†)	57.7
DeBERTa trained on GPT-4::aug (*long prompt*†)	**63.5** (2.5)
human classifier	89.4

Table 9: Correctly classified challenging sentences from the Twitter dataset (†: see Table 4).

classification scores.[10] This is proof that previous classifiers fail to detect our novel subtype of abuse. The classifier that performs best, i.e. GPT-4::aug, is also the classifier that was most successful in the detection of norm deviation (Table 5).

6.2 Experiments on the Twitter Dataset

We now evaluate on the Twitter dataset (§3.2) that comprises attested sentences. We use the same classifiers as in §6.1. For GPT-4::aug, we train on our constructed norm-compliance dataset rather than the Twitter dataset since we want to prove that our constructed dataset generalizes to realistic data.

Table 8 shows the results. Due to space limitations, the table only shows the scores on the entire dataset, i.e. we conflate all subtypes of abuse (*sexism*, *racism* etc.) to one single class. Similar to our constructed dataset, GPT-4::aug performs best.

After reviewing the errors made by our best classifiers, we identified a systematic error involving sentences that describe practices inherent to a specific identity group but uncommon among members of Western society (25)-(26). Even large language models, such as GPT-4, may misclassify these **challenging (non-abusive) sentences** as abusive instances. We manually identified 85 of such sentences in our Twitter dataset and computed the number of correctly classified sentences by our best classifier and also zero-shot classifiers using different language models. Table 9 shows the results. While there is still a considerable gap towards the human baseline, GPT-4 shows encouraging improvement over the other models.

(25) Muslims pray at dawn.
(26) Jews do not consume meat and dairy products together.

7 Conclusion

We addressed the task of detecting abusive sentences in which identity groups are depicted as deviating from the norm. We created novel datasets with sentences that do not express (explicitly) negative sentiment for this type of abuse via crowdsourcing. Previous classifiers are unable to detect this form of abuse sufficiently. This is a phenomenon not tied to specific lexical units. Therefore, only large language models produce good results, in our case, DeBERTa fine-tuned on data augmented by GPT-4. Our approach also handles negation and addresses non-abusive instances that are inherent to an identity group but not common for the Western society.

[10] The only exception is PerspectiveAPI in the detection of racism against Black people. Apparently, the training data for that tool appropriately represent all sorts of that type of abuse.

8 Limitations

We base our notion of what is norm-compliant and norm-contravening on **Western norms**.[11] This design choice was mainly driven by the availability of resources (i.e. language models and crowdworkers) which reflect the Western norm. We do not want to imply that the Western norm is more important than other norms. However, we believe that it is beyond the scope of a single research paper to duly address several norms at the same time.

Our dataset only addresses one subtype of abusive language. Therefore, classifiers trained on our new data are only capable to detect this subtype of abuse rather than abusive language, in general. Thus, we follow Wiegand et al. (2021b) who argue that a **divide-and-conquer approach** is the only reasonable approach to such complex phenomena. Ultimately, we envisage an array of different classifiers, each trained for a different subtype of abusive language (such as the task addressed in this paper) to be necessary to have a system that exhaustively detects abusive language.

In our data, we also observed cases in which a **norm-compliant** (rather than a norm-contravening) **property ascribed to an identity group is perceived as abusive**:

(27) Women usually work in an office.
(28) Women usually prepare the food.

Our classification approach is unable to detect such instances since they do not fall within its specialization. (27)-(28) are simply another type of implicit abuse, i.e. commonly observed stereotypes. We observed this phenomenon more frequently in the Twitter dataset than in the constructed dataset which is quite plausible as the latter will inevitably contain fewer stereotypes due to its creation process.[12] The absence of stereotypes in the constructed dataset also explains why the *correspondence* between deviating from the norm and abusive language is higher in the constructed dataset (Table 3) than in the Twitter dataset (Table 1).

Our research does **not target individuals**. Future work should investigate the extent to which the insights gained by the research presented in this paper are relevant to individuals.

[11]We did not explicitly enumerate the features of that norm to our crowdworkers, e.g. rule of law, pluralism, secularization, capitalism etc. Given that all crowdworkers were English native speakers living in Western countries and the strong consistency of their responses, we thought this was not necessary.

[12]The crowdworkers invented sentences addressing an otherwise unspecified group.

9 Ethical Considerations

Although our work clearly suggests that depicting a group of people as having properties or displaying a behaviour that deviates from the norm is often perceived as abusive language, **on no account do we want to imply that deviating from the norm is inherently reprehensible**. We would like to emphasize that the norm which is represented by our data (i.e. *Western norm*) may inherently reflect a bias towards heteronormativity. Furthermore, as many of our examples show, oftentimes the attribution of a behavior or property to a group is actually not warranted and instead is either an overgeneralization or a completely absurd claim. Thus, it is usually not the actual property or behaviour that makes people feel offended but the fact that this behaviour or property does actually not apply to them. This is supported by the fact that those behaviours and properties can actually be positive (29)-(30).

(29) Jews complete a Rubik's cube in under 10 seconds.
(30) Women usually know the names of at least 20 of their neighbours.

It is **not our intention to amplify the Western norm** with this research either. The classifiers we proposed are not designed to suppress certain properties or behaviours that deviate from the norm, in general. If inherent to the given identity group (25)-(26) they are not considered abusive and we specifically addressed these cases in our research.

One may **argue that any sentence targeting identity groups in general is abusive**, i.e. not only stereotypes or norm-contravention but also sentences such as *Jews use the internet*. While overgeneralizations are indeed problematic, our crowdworkers, who were members of the targeted identity group and judged whether a sentence was abusive, did not confirm this hypothesis.

Most of our new gold standard data were created with the help of crowdsourcing. All crowdworkers were compensated following the wage recommended by the crowdsourcing platform Prolific (i.e. $12 per hour). We inserted a warning of the offensive nature in the task advertisement.

In this work, we have crowdworkers create textual data representing abusive language since there is no alternative method that would yield a dataset with a comparable size and quality. In plagiarism detection (Potthast et al., 2010), deception detection (Ott et al., 2011) and abusive language detection itself (Vidgen et al., 2021b; Wiegand et al., 2021a) a procedure similar to ours was pursued.

10 Acknowledgements

The authors were partially supported by the Austrian Science Fund (FWF): P 35467-G. The authors would like to thank Sybille Sornig for contributing to the manual annotation of this research, and Julia Pardatscher and Ines Rehbein for their feedback on earlier drafts of this paper.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 54–59, Minneapolis, MN, USA.

Mohammad Ali and Naeemul Hassan. 2022. A Survey of Computational Framing Analysis Approaches. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9335–9348, Abu Dhabi, United Arab Emirates.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L. Williams. 2019. "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web*, 13(3):1–26.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 759–760, Perth, Australia.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of Association for Computational Linguistics: EMNLP 2020*, 1644–1650, Online.

Maria Becker, Katharina Korfhage, and Anette Frank. 2021. COCO-EX: A Tool for Linking Concepts from Texts to ConceptNet. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, pages 119–126, Online.

Joshua W. Buckholtz and René Marois. 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, 15:655–661.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 120–125, Sofia, Bulgaria.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment Propagation via Implicature Constraints. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 377–385, Gothenburg, Sweden.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4171–4186, Minneapolis, MN, USA.

Haibo Ding and Ellen Riloff. 2018. Weakly Supervsied Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 5763–5770, New Orleans, LA, USA.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 345–363, Online and Punta Cana, Dominican Republic.

Jon Elster. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal Communication*, 43(4):51–58.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Ernst Fehr and Urs Fischbacher. 2004. Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco,

Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1320, Online.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics (TACL)*, 10:178–206.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3309–3326, Dublin, Ireland.

Pengcheng Hea, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, Vancouver, Canada.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Mark R. Leary. 2000. Affect, cognition and the social emotions. In J. P. Forgas, editor, *Feeling and thinking: The role of affect in social cognition*. Cambridge University Press.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-Informed Transformations (LIT): A Method for Automatically Generating Contrast Sets. In *Proceedings of the BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online.

Björn Lindström, Simon Jangard, Ida Selbing, and Andreas Olsson. 2017. The Role of a "Common Is Moral" Heuristic in the Stability and Change of Moral Norms. *Journal of Experimental Psychology: General*, 147:228–242.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. 2014. A Theory of Blame. *Psychological Inquiry*, 25(2):147–186.

Julia Mendelsohn, David Jurgens, and Ceren Budak. 2021. Modeling Framing in Immigration Discourse on Social Media. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, 2219–2263, Online.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 1989–2005, Dubrovnik, Croatia.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim,

Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309–319, Portland, OR, USA.

Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 252–265, Toronto, Canada.

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

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 997–1005, Beijing, China.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 3027–3040, Seattle, WA, USA.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–58, Online.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. SOCIAL BIAS FRAMES: Reasoning about Social and Power Implications of Language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5477–5490, Online.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually Augmented Data and Unintened Bias: The Case of Sexism and Hate Speech Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4677–4695, Seattle, WA, USA.

Ravsimar Sodhi, Kartikey Panta, and Radhika Mamidi. 2021. Jibes & Delights: A Dataset of Targeted Insults and Compliments to Tackle Online Abuse. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH)*, pages 132–139, Online.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, volume 31, pages 4444–4451, San Friancisco, CA, USA.

June Tangney and Ronda Dearing. 2002. *Shame and Guilt*. Guilford Press.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal,

A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 33–42, Brussels, Belgium.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data. *PLoS One*, 15(12).

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 2289–2303, Online.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1667–1682, Online.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop*, pages 88–93, San Diego, CA, USA.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 5600–5612, Seattle, WA, USA.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 358–368, Online.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16280–16297, Singapore.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 576–587, Online.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.

Fabian Winter and Nan Zhang. 2018. Social norm enforcement in ethically diverse communities. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 115(11):2722–2727.

Tobias Wolbring, Christiane Bozoyan, and Dominik Langner. 2013. „Links gehen, rechts stehen!“ / "Walk Left, Stand Right!" Ein Feldexperiment zur Durchsetzung informeller Normen auf Rolltreppen / A Field Experiment on the Enforcement of Informal Norms on Escalators. *Zeitschrift für Soziologie*, 42(3):239–258.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double Perturbation: On the Robustness of Robustness and Counterfactual Bias Evaluation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 3899–3916, Online.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3952–3958, New York City, NY, USA.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit Sentiment Analysis with Event-Centered Text Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6884–6893, Online and Punta Cana, Dominican Republic.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):109:1–109:40.

Appendix Overview

This appendix provides more detailed information regarding certain aspects of our research for which there was not sufficient space in the main paper.

A Hyperparameters of Statistical Models

For all statistical models we used in this research we **refrained from heavy tuning of hyperparameters**. This is due to the fact that several experiments were evaluated in a cross-dataset setting, i.e. the training and test data originated from different datasets. As a consequence, tuning hyperparameters would only be possible by using some development data from the source domain. This, however, would mean that the resulting models would be tuned for the wrong domain. By running the tools with frequently used (default) settings of hyperparameters, we hope to produce models that are overall more robust across different domains (i.e. different datasets) than models fine-tuned on the wrong domain. Thus, we follow the strategy that was proposed for the large-scale cross-dataset evaluation reported in Wiegand et al. (2022).

A.1 Computing Infrastructure and Running Time

Our experiments were carried out on two servers:

- server 1: Lenovo ThinkSystem SR665; 1TB RAM; 2x32 Core AMD CPU that is equipped with 1 GPU (NVIDIA RTX A40, 48GB RAM)
- server 2: Quanton CS-221G-TRAN10-G12; 256GB RAM; Intel Xeon Silver 4310 that is equipped with 2 GPUs (both: NVIDIA RTX A40, 48GB RAM)

We estimate a total computational budget of 150 GPU hours.

A.2 PerspectiveAPI

In our evaluation, we also included *PerspectiveAPI*[13] as one baseline. This tool runs on unrestricted text and, from the publicly available classifiers, it is currently considered the state of the art for the general detection of abusive language (Röttger et al., 2021). The tool predicts several subtypes of abusive language. We considered the category *Identity attack* for our experiments, since it bears the greatest similarity concept-wise to the abusive sentences in our dataset.

[13] www.perspectiveapi.com

A.3 Transformers: BERT and DeBERTa

For classification, we fine-tuned both transformers BERT and DeBERTa using the implementation for text classification within the FLAIR framework (version 12) (Akbik et al., 2019). Regarding the exact models, we used `bert-base-uncased` and `deberta-large`. The former was chosen since of the currently available set of transformers, it is fairly small in size (110 million parameters) and thus has moderate resource requirements. The latter was chosen since it belongs to the more recent models of larger size (11.5 billion parameters) with notably improved performance but also greater resource requirements.

In order **not to overfit the model**, we chose the hyperparameter settings from Wiegand et al. (2022):

- learning rate=3e-5
- mini batch size=16
- maximal epochs=5

That work addressed similar data as our work (i.e. implicitly abusive language on identity groups). Since that work utilized RoBERTa (Liu et al., 2019) as its transformer model, we hope that those hyperparameters are not biased towards either BERT or DeBERTa.

A.4 GPT-4

We considered the most recent and persistent large language model that was available during the time span our experiments, i.e. GPT-4. In order to avoid overfitting, we mainly used the default settings of the hyperparameters:

- temperature=1.0
- top_p=1
- frequency_penalty=0.0
- presence_penalty=0.0

The only parameter for which we chose a setting different to the default settings is the maximum number of tokens (*max_token*). Our aim is to have a completion that comprises at least 2 sentences. Therefore, we applied the following setting:

- max_token=60

The reason for having two sentences instead of one sentence in the completion is as follows. In many instances, unlike in our previous examples (20)-(21), the first sentence of the completion is

insufficiently conclusive. For example, in (31) and (32), it is unclear what exactly GPT-4 considers norm-contravening, i.e. the negated proposition of the sentence to classify or the (plain) proposition without negation. As pointed out in §5.3, GPT-4 is inconsistent in that respect and may arbitrarily choose one of these scopes. The second sentence of the completion may help us to reconstruct what scope GPT-4 actually applied in its response. In (31), we thus learn that GPT-4 considers it uncommon to *not use the internet* (i.e. it considered the negated proposition), while in (32), GPT-4 considers *carrying around a bag of glitter to sprinkle when happy* to be uncommon (i.e. it considered the proposition without the negation).

(31) ***prompt:***
[Usually, they do not use the internet.]*sentence to classify*
Is this common in our Western society?
completion (first sentence):
No, this is not common in our Western society.
completion (second sentence):
The majority of people in the Western world use the internet on a regular basis for various purposes such as work, education, communication, entertainment, shopping, and so on.

(32) ***prompt:***
[They do not keep a large bag of glitter with them so they can sprinkle it when they feel happy.]*sentence to classify*
Is this common in our Western society?
completion (first sentence):
No, this is not common in Western society.
completion (second sentence):
Despite individual differences, carrying a bag of glitter to sprinkle when feeling happy is not a universally recognized or widely practiced behavior.

A.5 GPT-3.5

For GPT-3.5, we used the same settings as for GPT-4. Therefore, we refer the reader to Appendix A.4.

A.6 LLaMA-2

From LLaMA-2, we used the most advanced model available at the time of our experiments, specifically the `llama-2-70b-chat` model. This model was accessed via the API provided by Replicate.[14] We primarily used the model with the default hyperparameters as established by Replicate:

- temperature=0.5
- top_k=50
- top_p=1
- min_new_tokens=-1

Similar to our approach with GPT-4, we adjusted only the *max_tokens* hyperparameter. This adjustment was made for the same reasons as with the other language models (§A.4):

- max_tokens=60

[14] https://replicate.com

A.7 Logistic Regression

For logistic regression, we used the implementation within **LIBLINEAR** (Fan et al., 2008) **with L1 regularization**. The advantage of logistic regression is that it is a robust classifier which does **not require any hyperparameter tuning**.

B Details on Annotation

B.1 General Remarks

All **guidelines** for the different annotation and sentence generation tasks are contained in the supplementary material to this work, which is available upon request.

Our annotation guidelines as to what constitutes anti-Semitism, homophobia, Islamophobia, sexism, and racism were based on examples provided to us by members of the affected identity groups in an earlier crowdsourcing survey. We also asked our crowdworkers to rely on their intuition.

B.2 Details on Ranking Different Types of Implicit Abuse

In §1, we briefly mentioned that we also established via a crowdsourcing experiment that the phenomenon of implicitly abusive language we study in this paper is considered more severe than abusive comparisons and euphemistic abuse. For the sake of completeness, in the following we describe the set up of this elicitation experiment:

We asked crowdworkers to decide which of a pair of instances of implicitly abusive language they considered more severe. Each pair consisted of two different types of implicit abuse. The specific type of abuse was not revealed to the crowdworkers. We considered 6 other types of implicitly abusive language from existing datasets in addition to the form of abusive language we introduced in this paper. For each combination of types we had 20 different sentences (randomly sampled) rated by 5 crowdworkers each. Table 10 shows for each type the percentage it was considered more, less or equally abusive than the type it was paired with. The final ranking in Table 11 was computed based on the proportion a particular type of implicit abuse was rated to be more severe than the other type.

In this experiment, we sampled sentences of the data we produced as part of this research (§3) to serve as examples for the target phenomenon introduced in this paper. For the other forms of implicit abuse we selected sentences from other existing datasets. We specifically targeted those similar to

our newly identified type of abuse, i.e. subtly negative towards the target, and those explicitly labeled as such within these datasets. (Some types of implicitly abusive language mentioned in §2, such as *call for action*, lack a dataset in which this particular type is specifically marked as such.) Thus, we could avoid additional manual annotation. We use the following types:

Comparisons. This form of implicitly abusive language uses *like*-comparisons (33) from the dataset introduced by Wiegand et al. (2021a).

(33) You run like a headless chicken.

Euphemistic Abuse. This type of implicit abuse corresponds to the type we utilized as one of our cross-dataset classifiers, as detailed in §5.1.

Jibes. We consider the posts from Sodhi et al. (2021) extracted from the subreddit `r/roastMe` as a proxy of jibes (34).

(34) Where is the other half of your eyebrows?

Sarcasm. This form of implicit abuse taken from ElSherief et al. (2021) also includes humor and satire next to sarcasm. All those utterances have in common that their purpose is to attack or demean a protected class or individual (35).

(35) you want to hear a joke? women's rights

Stereotypes. This type taken from ElSherief et al. (2021) includes next to stereotypes (36) also misinformation, like holocaust denial.

(36) Can someone tell the black people in Chicago to stop killing one another before it becomes Detroit?

White Grievance. This form of implicit abuse was again taken from ElSherief et al. (2021). The authors define it as *frustration over a minority group's perceived privilege and casting majority groups as the real victims of racism* (37).

(37) Black lives matter and white lives don't? Sounds racist.

Contrary to the other crowdsourcing tasks for abusive language detection in our research, we did not specifically recruit crowdworkers from particular identity groups. This decision was made because many of the sentences to be rated in this task do not target any specific identity group.

B.3 Details on the Relevance of Sentiment in Our Datasets

As stated in §3, our datasets, both the constructed dataset and the dataset comprising sentences from Twitter, exclude sentences that convey explicitly negative sentiment. We excluded sentences conveying explicitly negative sentiment since existing datasets on abusive language detection involving identity groups convey overwhelmingly such negative sentiment (Wiegand et al., 2022). Therefore, classifiers trained on them will be able to cope with that type of sentiment.

By implicit sentiment, i.e. the sentiment type that remains in our dataset, we understand sentiment that is not conveyed by words with an unambiguously negative connotation, e.g. *poor*, *sad* or *bad*. Such words are also referred to as *sentiment* or *subjective expressions* (Wilson et al., 2005) that are also covered as part of sentiment lexicons, such as the *Subjectivity Lexicon* (Wilson et al., 2005). Though in general, such lexicons are a good approximation for establishing explicit sentiment automatically, we found that all publicly available lexicons are still sparse. Therefore, in order to provide even more accurate data, we refrained from simply using a lexicon look-up as an automatic procedure to identify explicit sentiment. Instead, we established it via manual annotation.[15] For this annotation, we also took into consideration the criterion of *defeasibility* (Deng and Wiebe, 2014). The relation between explicit/implicit sentiment and defeasibility is explained in the following:

(38) and (39) should be considered instances of implicit sentiment.

(38) They do not wash themselves very often.
(39) They do not work.

Apart from not containing any individual words conveying some unambiguously negative sentiment, the negative sentiment itself is defeasible: People may draw some negative conclusions from these sentences, e.g. (40) from (38) and (41) from (39), respectively. (Notice that (40) and (41) are instances of explicit sentiment.)

(40) They **smell unpleasantly**.
(41) They are **lazy**.

However, there are also other contexts possible that put these claims into a more neutral context, e.g. (42) for (38) and (43) for (39), respectively.

[15]This form of annotation was produced by one co-author of the paper who is a trained linguist.

	stereotypes	sarcasm	white_griev.	jibes	deviating_from_the_norm	comparisons	euphem.
is rated *more* abusive	58.2	54.0	45.9	41.6	34.2	26.8	22.2
is rated *less* abusive	20.5	25.1	34.2	39.6	47.7	54.9	60.8
is rated *equally* abusive	20.9	21.3	20.0	18.8	18.1	18.2	17.0

Table 10: Detailed results of crowdsourcing experiments to rank different forms of implicitly abusive language (*numbers represent percentages*).

most severe						***least severe***
stereotypes »	sarcasm »	white_grievance »	jibes »	**deviating_from_the_norm** »	comparisons »	euphemistic_abuse

Table 11: Ranking of different types of implicit abuse according to perceived severity.

(42) This is what their dermatologists recommend since these people are known to have very sensitive skin that also becomes inflamed fairly often.

(43) Instead they are studying for a higher degree at University.

On the other hand, with regard to instances of explicit sentiment (e.g. (40) and (41)), the sentiment is not defeasible.

To ensure that our definition of explicit sentiment is sufficiently well-defined, we also measured the inter-annotator between two co-authors. More details can be found in §B.4.

B.4 Details on the Inter-Annotator Agreement

For each of the variants of our constructed dataset (§3.1), we measured the inter-annotator agreement on a random sample of 200 sentences between one co-author and the majority vote of the labels provided by the crowdworkers. Table 12 (upper part) lists the agreement between these two annotations on each sample. It is highest on the norm-compliance task and lowest for *racism (Asians)*, though that agreement can still be considered substantial (Landis and Koch, 1977). The Twitter dataset (§3.2) underwent the same annotation tasks as the previously constructed dataset. We also used annotators from the same pool of crowdworkers.

For creating both the constructed dataset (§3.1) and the Twitter dataset (§3.2), it was necessary to remove or filter out content that contained explicitly negative sentiment. The notion of implicit and explicit sentiment as depicted in §B.3 was taken from Deng and Wiebe (2014). Since that annotation is much less dependent on the demographics of the annotators (for the subcategories of abusive language, we opted for the crowdworkers who are affected targets, e.g. *Jews*, *gay people*, *women* etc.) and due to financial reasons, this was done by one co-author, who is a trained linguist. However, we also measured the inter-annotator agreement to another co-author, who is also a trained linguist. The agreement was measured on a sample of 200 sentences invented by the crowdworkers that depict behaviours or properties as deviating from the norm (i.e. step ❶ in §3.1). The agreement (Table 12, lower part) was considered substantial (Landis and Koch, 1977).

feature	**Cohen's κ**
norm-compliance	0.86
anti-Semitism	0.85
homophobia	0.81
racism (Black people)	0.79
racism (Hispanics)	0.66
sexism	0.65
Islamophobia	0.64
racism (Asians)	0.62
explicitly vs. implicitly negative sentiment	0.83

Table 12: Inter-annotator agreement for the manual annotation of the dataset.

C Details on Debiasing the Norm-Compliance Dataset

After we had collected sentences displaying norm-contravening behaviours/properties and manually created norm-compliant counterparts ourselves, we inspected the resulting set of sentences for possible biases (Figure 1) by computing the Pointwise Mutual Information between words and each of the two classes of our dataset. There were notably more biases on the class *norm-compliant* than on the class *norm-contravening*. The reasons for that might lie in the way the sentences of the different classes were produced. As stated in §3.1, following the concept of *contrast sets* (Gardner et al., 2020), sentences for the class *norm-compliant* were produced by considering the sentences of class *norm-contravening*, e.g. (12), and converting them to a sentence of class *norm-compliant* by only applying minimal changes, e.g. (13). In contrast, for the *norm-contravening* sentences, crowdworkers were

			percentage of word occurrences with class *norm-compliant*	
rank	**word**	**overall frequency**	**before debiasing**	**after debiasing**
1	rarely	61	95.1%	54.2%
2	usually	26	92.3%	49.6%
3	may	63	77.4%	51.4%
4	occasionally	22	77.3%	45.2%
5	and	62	74.2%	52.7%
6	or	46	68.8%	54.5%
7	do	359	62.8%	55.4%
8	different	24	62.5%	52.4%
9	not	389	62.3%	55.1%
10	time	31	61.3%	53.6%

Table 13: Illustration of the biased word distribution on the class *norm-compliant* and impact of debiasing (words are ranked according to their bias towards the class *norm-compliant*).

fairly free to invent their content.

Table 13 shows the 10 most highly ranked words according to their Pointwise Mutual Information with class *norm-compliant* before debiasing. The last two columns show the percentage of instances of a word with class *norm-compliant* before and after debiasing. A word can be considered fairly unbiased if the percentage is about 50% since our norm-compliance dataset with 2 classes has almost a balanced class distribution (c.f. Table 1). Table 13 shows that there was a notable bias towards the class *norm-compliant* among these words before debiasing. If we judge those words based on their proportion in the class *norm-compliant* after debiasing, we can consider them sufficiently debiased. Please note that it is impossible to have exactly the same class distribution for all words. This is because, after our manual debiasing process, the entire dataset underwent the validation step (Figure 1). During this step, sentences were removed if crowdworkers considered them to be improper English or if they failed to reach a majority label. In other words: the size of the dataset still changed after we applied debiasing.

In order to keep the debiasing step at a tractable level, we only took into account words with a high Pointwise Mutual Information if they were also frequently occurring in our dataset. Individual words with a frequency of 10 or lower would not greatly affect learning methods. For lower frequency words, classifiers generally assign lower weights since they are only rarely observed.

Most biased words were function words. Content words are less likely to cause a harmful effect as our dataset covers many areas of life (Table 2) which prevent the same content word from occurring frequently.