Beyond *Negative* Stereotypes – Non-Negative Abusive Utterances about Identity Groups and Their Semantic Variants

Tina Lommel
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
tina.lommel@aau.at

Elisabeth Eder
Department for
High Performance Computing
Vienna University of Technology
AT-1040 Vienna, Austria
elisabeth.eder@tuwien.ac.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

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

Abstract

We investigate a specific subtype of implicitly abusive language, focusing on non-negative sentences about identity groups (e.g. *Women make good cooks*). We introduce a novel dataset comprising such utterances. It not only profiles abusive sentences but also includes various semantic variants of the same characteristic attributed to an identity group, allowing us to systematically examine the impact of different degrees of generalization and perspective framing. Thus we demonstrate that specific variants significantly intensify the perception of abusiveness. By switching identity groups, we highlight how the characteristics often described in stereotypes are not inherently abusive. We also report on classification experiments.

1 Introduction

Abusive language online is any language that could offend, demean or marginalize another person, covering the full range of inappropriate content from profanities and obscene expressions to threats and severe insults (Kiritchenko et al., 2021). Abusive language detection has been studied under a plethora of names, such as detection of *flaming* (Spertus, 1997), *cyberbullying* (Dadvar et al., 2013) or *hate speech* (Djuric et al., 2015). It poses a significant challenge for social media platforms. There is a pressing need for tools that aid content moderators in their decision-making processes.

A large portion of what is considered abusive language is categorized as implicitly abusive language, i.e. language **not conveyed by unambiguously abusive words** (Waseem et al., 2017; ElSherief et al., 2021). Detecting such language automatically remains a complex issue (van Aken et al., 2018; Wiegand et al., 2021b; Ocampo et al., 2023):

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

Implicitly abusive language often targets **identity groups**, i.e. groups of people that share a specific characteristic and therefore have a sense of unity (e.g. *Jews*, *gay people*, *women* etc.). A common form of targeting such identity groups is by imposing negative stereotypes on them:

(3) Gay people are **mentally unstable**.
(4) Muslim men **rape** women.
(5) Women always react **hysterically**.

The linguistic mechanisms underlying such negative stereotypes have also been examined (Wiegand et al., 2022). However, there are abusive stereotypical utterances addressing the same identity groups that are **not negative in sentiment**:

(6) Women make good cooks.
(7) Muslims have many children.
(8) Jews work in finance.
(9) Black people do not get a sun burn.

Based on a manual inspection of a random sample of 200 stereotypes about identity groups in the benchmark dataset from Nadeem et al. (2021), almost 40% of the instances were found to not be negative, i.e. a **very substantial proportion**.

In this paper, we introduce a new dataset for this type of utterance which is manually annotated

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28102–28120

via crowdsourcing. A distinctive feature of this dataset is that it includes the same stereotype expressed through a series of predefined **semantic variants**. This enables us to investigate how semantic changes of a stereotype (10), such as universal quantification (11) or instantiation (12), can either intensify or diminish its perceived abusiveness.

(10) Gay men like Cher.
(11) All gay men like Cher.
(12) Peter is gay and he likes Cher.

Our dataset also includes non-abusive utterances describing characteristics prevalent in identity groups rather than stereotypes (13). By subjecting these utterances to the same semantic variants and obtaining ratings for them, we can measure the degree to which such contextual elements (e.g. universal quantification) are intrinsically considered abusive when they co-occur with identity groups.

(13) Muslims pray in the direction of Mecca.

Our dataset encompasses 5 distinct identity groups. By **switching the identity group among the different stereotypical utterances**, we can also explore to what extent the perceived abusiveness of the characteristic described by a stereotype is linked to a specific identity group.

The work most closely related to ours is Wiegand and Ruppenhofer (2024), which addresses the task of detecting non-negative abusive sentences depicting identity groups as deviating from the (social) norm (7). Our paper examines the broader range of non-negative stereotypes that extend beyond such norm-contravening utterances to include other stereotypes, such as positive sentiment (6) or norm-compliant statements (8). Manual inspection on our dataset, which has been sampled from various sources (c.f. §3), revealed that among the non-negative abusive stereotypes, only less than 30% of the instances are norm contraventions. Therefore, our work focuses on a much broader subset of implicitly abusive language.

Wiegand and Ruppenhofer (2024) also analyzed the severity of abusive norm-contravening utterances. We replicated this experiment on our broader dataset and established that the wider range of non-negative stereotypes maintains their perceived severity as abusive, comparable to the norm-contraventions. Non-negative stereotypes were judged more severe than abusive euphemisms or comparisons (Wiegand et al., 2021a, 2023).[1]

[1]The exact details can be found in Appendix D.

We frame our task as a **sentence-level binary classification problem** where the two categories to distinguish are *abusive* and *not abusive*. The sentences are considered in isolation. We demonstrate that various classifiers trained on previous datasets fail to correctly classify such abusive utterances. Furthermore, because our dataset includes semantic variants in a structured manner, we are able to specify which variants are particularly challenging for both machines and humans.

Our contributions are the following:

- We introduce the first comprehensive dataset of non-negative stereotypes that are abusive.
- We examine the impact of various semantic variants of the same utterance systematically.
- Based on our dataset, we provide a descriptive analysis of non-negative abusive sentences.
- We examine in how far such abusive utterances can be automatically detected.

All data and annotation guidelines created as part of this research are **publicly available**.[2]

2 Related Work

Much of previous work in NLP regarding stereotypes has focused on analyzing stereotypical biases in language models (Bolukbasi et al., 2016; Caliskan et al., 2017; Cheng et al., 2023) and proposing measures to mitigate them (Sun et al., 2019; Zmigrod et al., 2019). In these contexts, some datasets have been introduced, the most common being CrowsPairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), the dataset introduced by Pujari et al. (2022) and SeeGULL (Jha et al., 2023). While only few studies address stereotype detection as an intrinsic task (Cryan et al., 2020; Fraser et al., 2022; Liu, 2024), many consider stereotypes in the context of abusive language detection (Fersini et al., 2018; Breitfeller et al., 2019; Caselli et al., 2020; Sap et al., 2020).

There has been a large body of research for abusive language detection (Nobata et al., 2016; Badjatiya et al., 2017; Fortuna and Nunes, 2018) with recent research focusing on specific implicit subtypes, such as dehumanization (Mendelsohn et al., 2020), dogwhistles (Mendelsohn et al., 2023), multimodal abuse (Kiela et al., 2020), comparisons (Wiegand et al., 2021a), euphemistic abuse (Wiegand et al., 2023) or abuse towards identity groups

[2]https://github.com/miwieg/beyond_negative_stereotypes

type	example
atomic	Jews are successful.
implication	*If* you are Jewish, you are successful.
"always"	Jews are *always* successful.
"all"	*All* Jews are successful.
"many"	*Many* Jews are successful.
"some"	*Some* Jews are successful.
instantiation	*Adam is Jewish* and he is successful.
self-identification	*We* Jews are successful.
authoritative report	*A recent survey published in the journal 'Economy and Society' reported that* Jews are successful.

Table 1: The different semantic variants.

	Jews	Muslims	gay	women	Black P.
crowdworkers	65	86	73	104	64
indiv. rating tasks	24	23	18	21	16
Cohen's κ	0.64	0.67	0.63	0.66	0.66
total sentences	2595	2515	2064	2361	1669
atomic sentences	291	286	241	267	189
avg. sent. length*	8.0	8.5	8.6	7.8	9.3
potent. stereotyp.	1882	1792	1377	1701	916
prevalent char.	713	723	687	660	753
abusive sentences	1038	468	980	1128	508
not abusive sent.	1531	2015	1067	1210	1144

Table 2: Statistics on the final dataset (*: *in tokens*).

(Davani et al., 2022; Hartvigsen et al., 2022). Our work also contributes to this last subtype.

While there has been some research in abusive language detection examining the role of context, most studies define context primarily as a broader segment of discourse specific to certain social media platforms. This includes research on Wikipedia conversations (Pavlopoulos et al., 2020), discussion threads on Fox News (Gao and Huang, 2017), conversation threads on Reddit (Vidgen et al., 2021; Yu et al., 2022) and pairs of reply Tweets (İhtiyar et al., 2023). An exception is the work by Chiril et al. (2020) who adopt a more linguistic approach to context by examining how reported speech influences the perception of sexism. Our research differs in that we create our own data, following a more controlled setup rather than exclusively focusing on naturalistic data. This approach also facilitates a more systematic analysis of different context variants.

3 Data

Our dataset focuses on **5 identity groups** that are also common targets of abusive language, i.e. *Black people*, *gay people*, *Jews*, *Muslims* and *women*. Our aim was to collect non-negative sentences for each of these identity groups. The most obvious subset of such utterances are stereotypes. These were obtained by browsing the Web (e.g. checking webpages of equality advocacy groups or social media platforms) and also asking members of these identity groups to provide a list based on their experience. For the latter, we recruited crowdworkers from Prolific.[3] Notice that the **focus of this work is not on the detection of stereotypes**. This is partly because the definition of stereotypes, especially in NLP research, is fairly vague. The definition from psychology that *stereotypes are generalized beliefs about categories of people* (Colman, 2015) is adopted. Though our collected data comply with this definition we do not claim that all of our sentences are unanimously accepted as stereotypes. Therefore, for the sake of linguistic accuracy, we refer to these utterances as **potential stereotypes**.

After removing every sentence that contained obvious negative sentiment[4] the remaining potential stereotypes were normalized into simple, basic sentence structures (i.e. matrix clauses) by one co-author. Such utterances presenting the identity groups in the bare plural form are also typically found in generic sentences (Cohen, 1999). Henceforth, we refer to these sentences as **atomic sentences**. This normalization was essential since we aim to systematically examine how various contexts affect the perception of each stereotype. To achieve this, we devised **8 semantic variants** that each of these atomic sentences were converted to as shown in Table 1. Our assumption is that the specific variant has an impact on whether people consider an utterance abusive. Our choice of variants was primarily guided by the concrete examples of stereotypical sentences we originally encountered. By some variants, we vary the degree of generalization (*"all"* vs. *implication* vs. *"many"* vs. *"some"* vs. *instantiation*) while in others we vary the extent to which the author identifies with the given proposition (*authoritative report* vs. *self-identification*).

To ensure a level of authenticity, we assert that the observations conveyed by the authoritative report were published in a reputable journal relevant to the subject area of the sentence. We considered journals that are listed in the first quartile of the SCImago Journal Rank rankings.[5] For the contexts labeled as *instantiations*, we selected stereotypical

[3] www.prolific.com

[4] We looked for negative polar expressions (*bad*, *horrible* etc.) and negated positive polar expressions (e.g. *don't like*).

[5] www.scimagojr.com

sem. variant	Jews	Musl.	gay p.	wom.	Black p.	avg.
atomic	48.1	22.4	63.9	65.5	36.0	47.2
implication	66.2	29.7	75.9	77.5	56.1	60.9
"always"	75.6	30.0	86.9	85.5	50.3	65.4
"all"	76.1	34.1	81.7	82.2	56.0	66.0
"many"	19.1	13.8	25.1	27.7	18.7	20.7
"some"	8.3	4.8	5.1	3.9	8.6	6.1
instantiation	6.3	7.6	10.0	5.3	12.7	8.0
author. rep.	22.3	10.5	17.7	19.9	10.1	16.5
self-identif.	38.1	14.7	57.3	62.5	24.9	39.6

Table 3: Percentage of sentences labeled as **abusive**.

first names that correspond to the respective identity group being discussed e.g. *Ali* for Muslims.[6]

Our dataset also includes **non-abusive sentences** that are structurally similar to the potentially stereotypical sentences. This similarity not only allows us to examine whether the structural form of sentences mentioning identity groups is perceived as abusive per se, but also provides essential non-abusive examples required for training classifiers. For these non-abusive sentences, we include sentences that describe **prevalent characteristics** of the respective identity groups (14)-(15). These were primarily collected from educational platforms and encyclopedias. The characteristics obtained were also instantiated with the 8 semantic variants. For some groups, particularly *gay people*, it was challenging to find such statements. In these instances, we created sentences attributing a characteristic to the group that aligns with common social norms (16).

(14) Jews eat three festive meals during Shabbat.
(15) Women get their period about once a month.
(16) Gay people cook dinner in their kitchen.

The sentences targeting a specific identity group were **rated by (additional) crowdworkers who identify with that particular group**. Often, the members of the targeted identity groups are the most skilled in recognizing this type of abusive language (Pei and Jurgens, 2023). The crowdworkers were again recruited from Prolific. In addition, only native speakers of English were admitted. The sentences were presented to the crowdworkers in random order. Each sentence could be labeled as *abusive*, *not abusive* or *unknown*. The latter label was intended for cases where the crowdworker was either undecided or did not understand the sentence. The final label is the majority vote of the 5 crowdworkers. While instances with tied ratings were included in the dataset, we excluded these from our subsequent classification experiments.

[6]The full list is shown in the Appendix in Table 13.

We also had all our atomic sentences rated by crowdworkers since we cannot assume that all potential stereotypes are perceived as abusive, nor can we presume that all prevalent characteristics about identity groups are perceived as not abusive.

Table 2 provides some statistics on the final dataset. For each identity group, the instances were divided into smaller rating tasks, with each task comprising about 100-130 sentences to be rated. A considerable number of crowdworkers contributed to annotating our dataset, which may enhance its representativeness. Table 2 also clearly shows that abusive sentences and potential stereotypes do not necessarily coincide. We particularly observed that among the identity groups *Muslims* and *Black people*, many potential stereotypes, especially those that are very positive, were not considered abusive.

For each subset targeting one identity group, we also measured the inter-annotator agreement on a random sample of 200 sentences. We had another group of 5 crowdworkers identifying as the targeted identity group rate these sentences and measured the agreement between the majority votes of the 5 new ratings and the original ratings. The agreement was substantial (McHugh, 2012), i.e. Cohen's $\kappa \geq 0.63$ for any identity group.

Unlike more general datasets for abusive language detection, such as ToxiGen (Hartvigsen et al., 2022) or the dataset introduced by Founta et al. (2018), our dataset is considerably smaller. It is intentionally focused and comparable in scope and scale to other recent contributions, such as Wiegand and Ruppenhofer (2024). Despite its relatively small size, the dataset offers high-value annotations for a specific subset of implicitly abusive language that remains underrepresented in existing resources. Our aim is to contribute depth and precision rather than breadth, which we believe constitutes a meaningful and complementary addition to the field.

4 Descriptive Analysis of the Dataset

Table 3 displays the proportion of sentences rated as abusive across the different semantic variants for each identity group in our new dataset.[7]

If we compare the distribution across the semantic variants, we see very similar tendencies within all 5 identity groups. The **semantic variants notably affect the proportion of sentences deemed**

[7]Appendix G includes an extended table taking into consideration potential stereotypes and prevalent characteristics.

type	prompt
common	Is this common?
fact	Is this a fact?
formal_language	Is this formal language?
negative_connotation	Could this sentence be interpreted as having a negative connotation toward the given identity group?
positive	Does this sentence ascribe a positive characteristic to the given identity group?

Table 4: GPT-4 prompts for feature extraction.

abusive (atomic) sentences (*random precision: 47.7*)		
feature	**prec**	**freq**
AUTO::GPT-4::negative_connotation (Yes)	85.7	273
MANUAL::preferences	78.1	178
AUTO::GPT-4::common (No)	75.7	411
AUTO::GPT-4::formal_language (No)	74.8	485
AUTO::GPT-4::fact (No)	66.4	740
MANUAL::work/education	63.4	82
MANUAL::competence	59.7	124
MANUAL::outward_appearance	59.3	145
not abusive (atomic) sentences (*random precision: 52.3*)		
feature	**prec**	**freq**
MANUAL::religious_practice	84.9	219
MANUAL::cultural/historical_characteristics	81.3	32
AUTO::GPT-4::fact (Yes)	79.4	501
AUTO::more_than_10_words	70.1	97
AUTO::GPT-4::formal_language (Yes)	69.3	774
AUTO::GPT-4::is_this_common (Yes)	66.6	808
MANUAL::physiology/biology/medicine	65.1	146
AUTO::GPT-4::negative_connotation (No)	62.7	985

Table 5: The 8 linguistic features (both automatic and manual) most strongly correlated with the 2 classes.

abusive. This strongly suggests that the form of the sentence (i.e. the semantic variants) also shapes perceptions rather than the characteristics described alone. For the variants *implication*, *"always"* and *"all"*, there is an increase in abusive instances compared to the atomic sentences. These three groups share a tendency to semantically generalize heavily towards entire groups. The opposite is true for the variants *"many"*, *"some"*, *instantiation*, and *authoritative report*. The first three of this latter set explicitly do not address the entire identity groups which makes it plausible that many fewer instances are considered abusive. Authoritative reports are also often perceived as not abusive. We assume that the author of an authoritative report is often interpreted as not sharing the view reported.

Contrary to our expectations, *self-identification* still exhibits a high proportion of abusive language. It appears that even when members of the targeted identity group utter such sentences, other members of the same group may still be critical of them.

To gain a clearer understanding of what constitutes abusive and not abusive *atomic* utterances in our dataset, we annotated these sentences with a set of linguistic features. Each feature was annotated either manually (by 1 person only) or automatically. The manual features (**MANUAL**) focus on the semantic topic a sentence addresses, e.g. *religious practice*. Among our automatic features (**AUTO**) is one that measures the length of sentences. All other automatic features are subjective and would have required manual annotation by multiple annotators, preferably through crowdsourcing, to produce reliable labels. However, given the strength of large language models in such tasks that perform at least on a par with human annotators (Gilardi et al., 2023; Chiang and Lee, 2023), we had these features extracted with GPT-4. Completions for the respective prompts (Table 4) starting with *Yes* and *No* were considered as 2 separate features.

Table 5 illustrates the 8 features most strongly correlated with either of the classes. (Table 17 in the Appendix displays a more comprehensive list of feature correlations.) Although our dataset does not include any sentences displaying negative sentiment, there is still a substantial number of sentences in which the sentence conveys some *negative connotation* and many of them are considered abusive. For instance, (17) implies a stereotype that women are overly emotional, and (18) could insinuate that Jews flaunt their wealth. Additionally, sentences exemplifying *othering* (Burnap and Williams, 2016), whether by depicting the target as deviating from social norms (19) or attributing unusual (biological) characteristics to them (20), are typically also perceived as abusive.

(17) Women love talking about their feelings.
(18) Jews wear expensive jewelry.
(19) Muslims don't go to the beach.
(20) Black people have a lower sense of pain.

As far as topics are concerned, sentences addressing preferences of identity groups are often regarded as abusive. On the other hand, if the sentences refer to religious, cultural or historical characteristics, then this is typically not perceived as abusive language. Similarly, factual and formal language[8] tends to be associated with non-abusive communication. We also observed that longer sentences correlate with non-abusive language. We

[8]To avoid spurious correlations in our dataset (Ramponi and Tonelli, 2022), we aimed to minimize notable stylistic differences. The perceived differences in formality may result from the mere use of technical expressions in some sentences, such as those discussing biological/medical conditions.

property	%
change of class label: any direction	44.4
change of label w. direction: *abusive* → *not abusive*	32.9
change of label w. direction: *not abusive* → *abusive*	11.5
characteristics maintaining label across all id. groups	20.9
label agreement among pairs of identity groups	
of most similar pair: *Black people* & *Muslims*	77.9
of most dissimilar pair: *Muslims* & *women*	48.6

Table 6: Analysis of how switching the identity groups in the dataset changes the perception of abusiveness.

type	prompt
abusive$_{plain}$	Is that sentence offensive?
abusive$_{sensitive}$	Is that sentence possibly offensive? In case of the slightest suspicion of offensiveness, please answer with "Yes".

Table 7: Additional prompts for zero-shot classifiers.

hypothesize that the length of a statement allows for more nuanced content. In contrast, generalizations and oversimplifications, which form the basis of the stereotypical sentences, tend to be brief.

5 Switching Identity Groups

In order to find out in how far a certain characteristic described in a sentence is only perceived abusive in conjunction with a specific identity group, we **substituted the mention of the identity group by one of the other 4 groups**. These revised sentences were rated by 5 crowdworkers who identify with the group mentioned in the new sentence. In previous work that focused on utterances ascribing *negative* characteristics to an identity group, it was found that it usually does not matter whether the resulting sentence represents a well-known stereotype (21) or not (22) (Wiegand et al., 2022). In the present work, we assume that with non-negative characteristics the perception of abusiveness is more dependent on the identity group.

(21) Muslims are terrorists. *(abusive and well-known stereotype)*
(22) Women are terrorists. *(abusive but no well-known stereotype)*

The sentences provided to crowdworkers for a specific identity group were randomly mixed with sentences from our original dataset that targeted that identity group. These instances primarily served as distractors, as we had already obtained ratings for them from previous experiments (§3). To maintain manageability, this elicitation study was confined to the atomic sentences. Similar to our previous experiments (§3), we considered the majority vote of 5 ratings as the final class label.

Table 6 presents statistics for this additional dataset. Nearly half of the sentences receive a different class label when the identity group is substituted. Changes from abusive sentences to non-abusive ones occur more often than changes in the opposite direction. Obviously, many abusive sentences (23) cease to be viewed as stereotypes when the identity group is replaced (24).[9] Only few abusive sentences continue to reinforce the underlying stereotype regardless of the identity group mentioned (25).[10] Non-abusive utterances depicting common characteristics, such as (26), however, remain non-abusive to nearly every identity group.

(23) Jews work in finances. *(abusive)*
(24) Women work in finances. *(not abusive)*
(25) <*identity_group*> have the same jobs. *(abusive)*
(26) <*identity_group*> eat on a regular basis. *(not abusive)*

Table 6 also provides the percentage of label matches between the most similar and the most dissimilar pair of identity groups.[11] The proportion of matches varies significantly. This suggests that cross-group classification (i.e. a classification setup in which sentences from different identity groups are used in training than those comprising the test data) will only be effective to a certain extent; specifically, if the training data originate from identity groups that share similar stereotypes (e.g. perceptions of foreignness and strong community/family bonds for *Black people* and *Muslims*).

6 Classification Experiments

6.1 The Different Classifiers

For supervised learning, we use DeBERTa (Hea et al., 2021), which is one of the most sophisticated publicly available transformers. We fine-tune the pretrained model 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 more details on the settings of all classifiers.***

We examine the following classifiers:

Existing Classifiers. We consider 2 publicly available classifiers for abusive language detection in order to examine in how far they can detect the

[9] Though (24) might still be seen as an overgeneralization, it is presumably not considered abusive as it contrasts with the abusive stereotype of women as primarily housewives.
[10] Table 6 in the Appendix provides more such examples.
[11] Table 16 in the Appendix shows the scores for all pairs.

		best	worst
classifier	**average**	**atomic**	**"some"**
majority-class	37.3	34.3	48.4
variant_generalization	41.8	34.3	48.4
PerspectiveAPI	52.9	56.1	54.8
StereoSet*	53.2 (1.4)	54.2 (1.8)	50.0 (0.8)
ToxiGen	56.0	58.7	57.8
ISHate*	57.1 (1.9)	60.7 (2.3)	54.3 (1.4)
cross-group*	58.6 (1.0)	68.5 (0.5)	48.4 (0.1)
Pujari et al.*	61.0 (1.5)	66.2 (1.3)	55.6 (1.2)
atomic_classifier::auto*	62.5 (1.4)	68.3 (2.0)	51.4 (1.3)
GPT-4::zero::abusive$_{plain}$	63.9	65.9	57.5
within-group*	64.2 (2.1)	73.5 (0.4)	52.2 (4.3)
cross-group+within-group*	65.2 (1.6)	75.2 (1.1)	50.6 (4.4)
GPT-4::zero::fact	65.4	72.0	58.4
GPT-4::zero::formal	66.0	71.3	62.1
GPT-4::zero::neg._connot.	66.0	70.2	60.4
GPT-4::zero::common	65.3	68.9	56.7
GPT-4::zero::abusive$_{sensitive}$	67.4	72.7	63.4
GPT-4::augmenting*†	**70.2** (1.0)	**78.3** (0.9)	**62.0** (2.0)
atomic_classifier::oracle	73.1	100.0	59.6
human_classifier	76.9 (1.0)	82.1 (0.5)	69.8 (2.3)

Table 8: F1-scores of classifiers averaged over all semantic variants and the performance of the *best* and *worst* performing variant (i.e. *atomic* and *"some"*); numbers in brackets denote standard deviation; *: fine-tuned with DeBERTa; †: uses the completions of the 5 best zero-shot approaches (*prompts are shown in Tables 4 & 7*).

type of abusive language in our dataset. We use: ***PerspectiveAPI***,[12] 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***.

Implicit Abuse. Since we consider non-negative abusive sentences as instances of abusive language, we also fine-tune a transformer on the ***ISHate*** dataset (Ocampo et al., 2023), a dataset consolidating 7 existing datasets for implicit abuse.

Stereotype Classification. We fine-tune a classifier on each of 2 recent datasets for stereotype classification, i.e. a classifier that distinguishes between stereotypes and non-stereotypes. The prediction of a stereotype is considered as a proxy for abusive language. Thus, this baseline tells us how much the type of abusive language our dataset coincides with stereotypes. As datasets, we chose the well-known datasets ***StereoSet*** (Nadeem et al., 2021) and the dataset by ***Pujari et al. (2022)***.

Cross-Group Classification. We want to simulate how classifiers trained on our dataset would score on an unseen identity group since our dataset only comprises a small set of different identity groups. Therefore, we train classifiers on 4 identity groups and test on the remaining one.

Within-Group Classification. This classifier is trained on the instances of the same identity group using a 5-fold crossvalidation. The folds are arranged in such a way that similar topics are kept within the same fold. This enables a fairly strict evaluation in which the test fold always includes topics that have not been observed during training.

Variant Generalization. This classifier assigns the most frequently occurring class label to all instances of a specific variant. For example, all sentences in the *always*-variant are classified as abusive as this is the most frequent class observed with this variant (Table 3). This approach provides an estimate of how well a classifier performs by focusing solely on the semantic construction of the variant and disregarding the content of the utterance.

Atomic Classifier. This classifier has no knowledge of semantic variants and treats them as atomic sentences. Thus, we can estimate how much we lose by ignoring the subtleties in meaning caused by the semantic variants. We consider **2 versions**: The first, **oracle**, maps the actual label from the gold standard of an atomic sentence to all its semantic variants as predictions (so all atomic sentences are classified correctly). The second, **auto**, uses labels from a classifier trained solely on atomic sentences (but is tested on all sentences).

GPT-4. Given the high effectiveness of large language models (Laskar et al., 2023), we also consider GPT-4 as a classifier. Apart from re-using the prompts we used for the descriptive feature analysis in §4, we also asked for abusive language directly (Table 7). These 2 prompts differ in the degree of sensitivity. We use the term *offensive* in the prompts rather than *abusive*, as the language model responds more effectively to it, likely due to its more prevalent use in language data and, consequently, in the model's training data. We consider **2 types of classifiers** based on GPT-4:

Our first classifier is a **zero-shot** classifier (Plaza-del-arco et al., 2023) that uses one of our previous prompts (Tables 4 & 7). The classifier maps completions to a prompt beginning with *Yes* or *No* to the labels *abusive* and *not abusive*.[13]

In our second classifier, we **augment** each sentence in our dataset with the respective completions obtained from the zero-shot approach: we concate-

[12] https://perspectiveapi.com

[13] The exact mapping depends on the semantics of the prompt, e.g. for the type *common* (Table 4) we map *No* to *abusive* but for *negative connotation* we map *Yes* to that label.

	F1	percentage change relative to atomic variant								
classifier	**atomic**	**implic.**	**"always"**	**"all"**	**"many"**	**"some"**	**instant.**	**author.**	**self-id.**	**average**
majority-class	34.3	-19.8	-26.2	**-27.4**	+28.9	+41.1	+39.7	+32.7	+9.3	+8.8
variant_generalization	34.3	+11.7	+16.0	+16.9	+28.9	+41.1	+39.7	+32.7	+9.3	+21.9
StereoSet	54.2	-2.0	+4.8	-1.1	-4.6	**-7.8**	-4.2	+0.4	-1.7	-1.9
PerspectiveAPI	56.1	-6.8	-4.5	-5.7	-7.1	-2.3	-10.0	**-10.3**	-4.3	-5.7
ToxiGen	58.7	-6.5	**-14.1**	-8.7	-8.2	-1.5	-1.2	-4.3	-13.5	-4.6
ISHate	60.7	-5.6	-4.5	-2.1	-7.9	**-10.5**	-7.7	**-10.5**	-4.3	-5.9
GPT-4::zero::abusive$_{plain}$	65.9	+4.3	-4.3	+0.8	-1.4	**-12.8**	-2.6	-4.6	-7.3	-3.0
Pujari et al.	66.2	-4.1	-8.5	-1.4	-11.3	**-16.0**	-14.4	-13.1	-1.8	-7.9
atomic_classifier::auto	68.3	-1.2	+0.2	-1.3	-13.3	**-24.7**	-24.0	-13.2	+1.8	-8.5
cross-group	68.5	-8.0	-20.6	-14.0	-15.8	**-29.3**	-21.8	-14.9	-5.0	-14.5
GPT-4::zero::common	68.9	+8.0	-3.1	-0.2	-9.0	**-17.7**	-6.8	-10.9	-7.3	-5.2
GPT-4::zero::neg._connot.	70.2	-2.3	-3.0	-3.1	-7.8	**-14.0**	-7.7	-8.0	-7.7	-6.0
GPT-4::zero::formal	71.3	+3.5	-2.8	-1.1	-11.6	-12.9	-17.0	**-18.1**	-7.2	-7.4
GPT-4::zero::fact	72.0	+2.6	-12.8	-8.3	-11.0	**-18.9**	-18.5	-18.5	+2.1	-9.2
GPT-4::zero::abusive$_{sensitive}$	72.7	+2.6	-4.4	-3.9	-8.7	-12.8	-12.5	**-13.9**	-11.8	-7.3
within-group	73.5	-4.1	-3.5	-4.8	-23.4	**-29.0**	-26.9	-22.3	-0.41	-12.7
cross-group+within-group	75.2	-4.9	-5.7	-5.2	-23.0	**-32.7**	-26.2	-20.7	-1.9	-13.3
GPT-4::augmenting	78.3	-3.1	-3.8	-3.8	-17.6	-20.8	**-23.4**	-19.9	-0.8	-10.3
human_classifier	82.1	-2.4	-1.5	-2.4	-8.5	**-15.0**	-12.8	-11.6	-2.9	-6.3
atomic_classifier::oracle	100.0	-23.0	-27.7	-24.8	-30.8	**-40.4**	-39.6	-33.5	-22.5	-26.9

Table 9: Percentage change relative to F1-score of the *atomic* variant for all classifiers (highest decrease in **bold**).

nate to each sentence the completions from the 5 best zero-shot classifiers examined in this work. We then fine-tune and test a transformer on these augmented instances. Notice that the completion itself does not only comprise a mere *Yes* or *No* but sufficient text for text explaining the rationale behind the language model's response (c.f. Appendix A.4). This classifier combines plain within-group classification with zero-shot classification.

Human Classifier. As an upper bound, we tested a classifier in which we randomly sampled the judgment of one individual annotator from the crowdsourced gold-standard annotation.

6.2 Evaluation

Classifier Comparison. Table 8 displays the performance of the different classifiers, averaged over all semantic variants and on both the overall best and worst performing variants.[14]

The classifiers trained on existing datasets for abusive language or stereotype detection perform poorly. This can be attributed to the fact that these datasets do not fully address the linguistic phenomena targeted by our dataset. In particular, non-negative stereotypes are underrepresented, and the semantic variations we systematically examine are only sporadically included.

Regarding classifiers trained on our new dataset, there is a clear advantage to training on data focusing on the target identity group (i.e. *within-group*)

classifier	**Jews**	**Musl.**	**gay**	**wom.**	**Black**
majority-class	37.3	44.8	34.3	34.1	40.9
StereoSet	44.1	51.6	53.8	53.5	54.4
atomic_classifier::auto	58.6	56.7	62.1	55.7	49.3
PerspectiveAPI	59.2	59.9	56.5	60.4	60.7
Pujari et al.	67.9	63.4	64.8	67.3	57.7
ToxiGen	69.9	65.0	62.9	70.1	56.3
ISHate	71.0	60.8	63.1	68.1	62.5
GPT-4::zero::neg._con.	70.2	73.7	70.6	71.7	62.8
variant_generalization	74.0	61.7	74.0	74.3	68.4
GPT-4::zero::abus.$_{plain}$	72.3	75.0	72.7	68.0	65.9
GPT-4::zero::abus.$_{sensit.}$	77.9	76.6	71.5	75.7	61.7
GPT-4::zero::common	74.7	74.1	72.8	76.8	66.0
GPT-4::zero::fact	74.6	70.3	71.6	80.4	64.9
GPT-4::zero::formal	76.5	75.7	72.9	76.4	64.7
cross-group	74.5	68.4	77.8	80.5	67.5
within-group	76.9	71.7	80.9	81.9	64.1
cross-gr.+within-gr.	77.2	69.1	81.3	82.5	66.1
GPT-4::augmenting	**79.8**	**77.0**	**82.9**	**84.0**	**67.8**
atomic_classifier::oracle	69.0	77.7	67.0	67.4	64.4
human_classifier	81.4	80.2	82.4	84.4	73.6

Table 10: F1-score of classifiers on each identity group aggregated over all semantic variants.

[14]Table 18 shows the results on each individual variant.

compared to training on other identity groups. Classifiers generalizing across all sentences of a semantic variant and classifiers that disregard the semantic variant (i.e. *atomic_classifier*) do not perform well either. From this we conclude that the abusiveness arises both from the semantics of the atomic sentence and the context provided by the semantic variant.

The zero-shot classifiers achieve respectable performance. There is also a notable improvement when moving from $abusive_{plain}$ to $abusive_{sensitive}$. This suggests that a plain prompt may lead to overly conservative predictions of abusive language. The best overall classifier combines within-group training data augmented with GPT-4 completions.

Semantic Variants. Table 8 shows that atomic sentences are the easiest to classify, likely due to the absence of contextual embedding.

Table 9 presents the performance change relative to the atomic variant for the other semantic variants. Overall, we observe a trend: the better the overall performance on atomic sentences, the more pronounced the performance drop across the various semantic variants. The most challenging instances involve *authoritative report*, *instantiation* and quantifications using *"many"* and *"some"*. These variants also present the greatest challenges to human annotators. From a semantic standpoint, these sentences tend to generalize less toward the identity group compared to other variants. Some of them are still perceived as abusive if the utterance involves a very common stereotype:

(27) Mary is a woman and she knows how to iron.
(28) Some Jews have an innate business sense.

What makes the *"some"*-variant particularly challenging is that several abusive instances, e.g. (29), do not refer to a common stereotype (30), but instead imply a negative character trait (31).

(29) Some women are compliant.
(30) Women are compliant.
(31) Most women are uncooperative.

Identity Groups. In Table 10, the performance for each identity group is listed. All those scores are aggregated over all semantic variants. This table shows variability in performance across different identity groups, with *women* achieving the best results and *Black people* the worst. The lower performance for the latter group may be due to the limited amount of training data available for this identity group (see Table 2). It may also be related to a higher degree of disagreement among human annotators on this data, as reflected in the comparatively poor performance of the human classifier for this group.

We also observe that the gap between the best classifier (i.e. *GPT-4::augmenting*) and the human classifier (particularly *gay people* and *women*) is notably smaller than that observed in Table 8, where performance is evaluated on the basis of semantic variants. From that we conclude that the evaluation focusing on identity groups (Table 10) is too optimistic while a focus on semantic variants (Tables 8 and 9) highlights limitations even in state-of-the-art classifiers.

7 Conclusion

We introduced a dataset focusing on non-negative sentences about identity groups for abusive language detection. We demonstrated that there is a complex interplay between the mention of a specific identity group, the characteristic attributed to them and a particular semantic variant. By varying the same proposition across different semantic variants, we could show that the semantic variant has a notable impact on the perception of the sentence. Specifically, variants that generalize a characteristic are perceived as markedly more abusive. By switching identity groups, we showed that many characteristics are only perceived abusive with a mention of some particular identity group. Classifiers trained on previous datasets are unable to cope with these data satisfactorily. A classifier using GPT-4 completions produced best performance.

8 Limitations

Our dataset is constructed and may be criticized for its **artificial nature**. However, such a constructed dataset is the only means through which we can systematically examine different semantic variants. Moreover, while our data are not attested or naturalistic, they are elicited, which differs from being purely *synthetic*. Our elicited data are grounded in real-world intuitions, making them less artificial than fully synthetic data produced by automated means or without human judgment.

We **analyzed sentences in isolation**. Although, ultimately, we aim to address utterances in a broader context (Bourgeade et al., 2024), we believe it is essential first to understand the mechanisms underlying a more basic context. Our experiments also show that already the isolated sentences

present a challenge for previous classifiers.

We **focus on 8 semantic variants**. While further variants could influence the perception of abusiveness, we selected those that we typically observed in our data (§3). Also, the current choice has a significant impact because they either vary the degree of generalization or involve perspective framing. We excluded singular definite generics (32) as they were not part of our observed data and are known to be associated with pejorative connotations (Palmer et al., 2017), potentially biasing their use.

(32) The Jew is successful.

We also concentrated on aspects that can be easily adapted to our sentences framework. Other linguistic dimensions are known to impact the perception of abusiveness, such as aspect (Friedrich and Pinkal, 2015). For instance, episodic aspect (33) is perceived as less abusive (Wiegand et al., 2022). However, converting our atomic sentences, which are habitual, into episodic ones is not straightforward and may even be impossible, as kinds are typically not involved in episodic or individual events.

(33) Muslim assassinates 2 Christian aid workers.

Our work **addresses only 5 identity groups**, despite the existence of many others. These 5 identity groups are among those most commonly targeted in English-language discourse in Western contexts. We are confident, however, that the insights gained are broadly representative of the phenomenon, as the selected groups reflect diverse dimensions such as gender, religion, sexuality and ethnicity, that is, categories that are frequently the target of abusive language. While it would have been ideal to include a wider range of identity groups, this was beyond the financial and practical scope of the present study.

In our research, we primarily concentrated on GPT-4, a large language model. One limitation of GPT-4 is that it is proprietary software. Initially, we also evaluated open-weight language models, including LLaMA-2, but found their performance significantly lacking in comparison. Since our study is intended as a proof of concept, it was essential to employ the most advanced methodologies available to demonstrate the highest level of performance currently achievable in this field.

In this study, we found that specific styles of language, such as formal language and lengthy texts, are generally perceived as non-abusive. This finding does not necessarily imply that abusive language, particularly non-negative abusive utterances, is absent in such texts. Still, to most people, these texts typically appear less abusive.

In our classification experiments, we discovered that on certain semantic variants, most notably on *"some"* and *instantiation*, all classifiers perform poorly. This should **not** be viewed as an indication that our proposed classifiers are insufficiently mature. We also observed that the performance with our human classifier similarly dropped. This suggests that these are semantic variants generally recognized as extremely challenging. It also highlights which semantic contexts of abusive language are those on which people are most divided.

9 Ethical Considerations

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.

We did not ask crowdworkers to *invent* stereotypes. Instead, we asked them to *report* stereotypes they had personally encountered and considered problematic. We also ensured that the crowdworkers understood this information was being collected solely for linguistic research purposes and emphasized that all researchers involved in the study were aware of the sensitive nature of such utterances. Additionally, the legal department of our institution was informed about the nature of this research.

While the prevalent characteristics we refer to are common among members of the relevant identity groups, we recognize that they do not necessarily apply to all individuals within those groups. However, these characteristics differ from potential stereotypes, as they are grounded in biological, medical or scientific reality, rather than being oversimplified notions that reflect social or cultural expectations instead of objective facts. This also explains why these prevalent characteristics are less likely to be perceived as abusive language.

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.

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.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 4356–4364.

Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixin Sun. 2024. Humans Need Context, What about Machines? Investigating Conversational Context in Abusive Language Detection. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8438–8452, Torino, Italia.

Luke M. Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1664–1674, Hong Kong, China.

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 6193–6202, Online.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1504–1532, Toronto, Canada.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15607–15631, Toronto, Canada.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4055–4066, Online.

Ariel Cohen. 1999. *Think Generic!: The Meaning and Use of Generic Sentences*. CSLI, Stanford.

Andrew M. Colman. 2015. *A Dictionary of Psychology*. Oxford University Press.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–11, Honolulu, HI, USA.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Proceedings of the European Conference in Information Retrieval (ECIR)*, pages 693–696, Moscow, Russia.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2022. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics (TACL)*, 11:300–319.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the International Conference on World Wide Web (WWW Companion)*, pages 29–30.

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.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*.

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.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Stanford, CA, USA.

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2022. Computational Modeling of Stereotype Content in Text. *Frontiers in Artificial Intelligence*, 5(826207).

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2481, Lisbon, Portugal.

Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 260–266, Varna, Bulgaria.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

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*.

Musa İhtiyar, Ömer Özdemir, Mustafa Erengül, and Arzucan Özgür. 2023. A Dataset for Investigating the Impact of Context for Offensive Language Detection in Tweets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1543–1549, Singapore.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9851–9870, Toronto, Canada.

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.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada.

Yang Liu. 2024. Quantifying Stereotypes in Language. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1223–1240, St. Julian's, Malta.

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*.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Julia Mendelsohn, Rona Le Bras, Yejin Choi, and Maarten Sap. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15162–15180, Toronto, Canada.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5356–5371, Online.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online.

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.

Alexis Palmer, Melissa Robinson, and Kristy K. Phillips. 2017. Illegal is not a Noun: Linguistic Form for Detection of Pejorative Nominalizations. In *Proceedings of the Workshop on Abusive Language Online*, pages 91–100, Vancouver, BC, Canada.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305, Online.

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.

Flor Miriam Plaza-del-arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In *Proceedings of the Workshop on Online Abuase and Harms (WOAH)*, pages 60–68, Toronto, Canada.

Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. 2022. Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6703–6712, Dublin, Ireland.

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.

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.

Ellen Spertus. 1997. Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of Innovation Applications in Artificial Intelligence (IAAI)*, pages 1058 – 1065.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1630–1640, Florence, Italy.

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, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. 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.

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.

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 and Josef Ruppenhofer. 2024. Oddballs and Misfits: Detecting Implicit Abuse in Which Identity Groups are Depicted as Deviating from the Norm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2200–2218, Miami, FL, USA.

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.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 5918–5930, Seattle, WA, USA.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1651–1661, Florence, Italy.

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 one GPU (NVIDIA RTX A40, 48GB RAM)
- server 2: Quanton CS-221G-TRAN10-G12; 256GB RAM; 1 Intel Xeon Silver 4310 that is equipped with two GPUs (both: NVIDIA RTX A40, 48GB RAM)

We estimate a total computational budget of 80 GPU hours.

A.2 PerspectiveAPI

In our evaluation, we also included *PerspectiveAPI*[15] 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.

A.3 DeBERTa

For classification, we fine-tuned DeBERTa (more specifically `deberta-large`) using the implementation for text classification within the FLAIR framework (version 12) (Akbik et al., 2019). This language model belongs to the more recent models of larger size (11.5 billion parameters). It has been reported to achieve high classification performance in different NLP tasks.

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 detection). Since that work utilized RoBERTa (Liu et al., 2019) as its transformer model, we hope that those hyperparameters are not biased towards DeBERTa.

A.4 GPT-4

We mainly used the default settings of the hyperparameters of GPT-4:

- temperature=1
- top_p=1
- frequency_penalty=0
- presence_penalty=0

The only parameter for which we chose a setting different to the default settings in all of our experiments is the maximum number of tokens (*max_tokens*). Our aim was to have a completion that is sufficiently long to answer the questions of our prompts. In our exploratory experiments,

[15] www.perspectiveapi.com

we determined that the following parameter setting achieved that goal:

- max_tokens=60

Please note that we only conducted a limited number of experiments, varying the parameter values at coarse-grained intervals. We generally selected a value that typically elicited responses covering **two complete sentences** to address the prompt. Two sentences were necessary because while the first sentence typically answers the prompt's question, only the second provides a more substantial explanation (34). We did not explicitly optimize the aforementioned token parameter for classification performance, as doing so would have caused overfitting to our dataset. Moreover, retaining this parameter at its default setting (i.e. *max_tokens*=256) was not considered an option either, since such responses would have been unnecessarily verbose, leading to undue financial costs.

(34) **sentence to classify:** Women prefer playing softer sports.
prompt: Could this sentence be interpreted as having a negative connotation toward the given identity group? Please answer with "Yes" or "No". In addition, please briefly motivate your answer.
completion: [Yes, the sentence could be interpreted as having a negative connotation toward the given identity group.]$_{first\ sentence}$ [It could be seen as reinforcing stereotypes about women being less capable or less interested in more physically demanding or aggressive sports, which might not be true for all women.]$_{second\ sentence}$

B StereoSet

Unlike the other datasets we used for our experiments, building a stereotype classifier using StereoSet (Nadeem et al., 2021) involved several degrees of freedom. Therefore, we describe here the steps we undertook.

StereoSet was created via crowdsourcing. Crowdworkers were asked to provide associations for a set of given identity groups. The associations were specified in two different tasks:

Intrasentence task. An incomplete sentence is provided to the crowdworkers (e.g. *Girls tend to be more ___ than boys*), who are then required to complete it.

Intersentence task. Crowdworkers receive a complete sentence (e.g. *He is an Arab from the Middle East*) and are asked to provide associative sentences.

For both tasks, the resulting texts were labeled into one of three categories: *stereotypes*, *anti-stereotypes* and *meaningless*. To build a binary classifier that distinguishes stereotypes from other types of sentences, we included instances from both tasks. For the intrasentence task, we considered the completed sentence as an instance. For the inter-sentence task, we considered the concatenation of the given sentence and the associative sentence as an instance.

We trained a classifier on two categories: *stereotypes* (as a proxy for abusive language) and *anti-stereotypes* (as a proxy for non-abusive language). We did not include instances labeled as *meaningless* since they often resulted in nonsensical sentences.

C How Crowdworkers were Recruited

All crowdworkers were required to be **native speakers of English**. Additionally, each crowdworker had to identify with one of the 5 identity groups we consider in this paper. These identity groups were available as screening options on Prolific. To ensure reasonable annotation quality, we permitted only those crowdworkers who had an approval rate of 100% and certified they did not have dyslexia to participate in our surveys.

Our annotation tasks were divided into smaller segments for the crowdworkers, such as evaluating the labels of 130 sentences. This approach was designed not only to alleviate their annotation burden but also to enable a larger number of members from the targeted identity group to participate (approximately 80 crowdworkers per identity group on average, cf. Table 2), thereby obtaining a more representative rating to constitute our final gold standard.

D Experiment on Perceived Severity of Non-Negative Stereotypes

As mentioned in §1, following Wiegand and Ruppenhofer (2024) we had the non-negative abusive stereotypes be rated according to the perceived abusiveness. More specifically, we also had crowdworkers, all native English speakers without specific backgrounds, compare examples of non-negative abusive stereotypes of our dataset with other types of implicit abuse according to their perceived severity. The examples (20 instances for each type) were presented in pairs without revealing their types. Crowdworkers had to decide which

	sarcasm	**white_griev.**	**jibes**	**non-negative_stereotypes**	**comparisons**	**euphem.**
is rated *more* abusive	60.7	60.4	43.3	34.9	25.3	21.0
is rated *less* abusive	20.4	20.7	36.7	46.8	56.8	62.9
is rated *equally* abusive	18.8	18.4	20.0	18.2	17.8	16.2

Table 11: Detailed results of crowdsourcing experiments to rank different forms of implicitly abusive language (*numbers represent percentages*).

most severe	***least severe***
sarcasm » white_grievance » jibes » **non-negative_stereotypes** » comparisons » euphemistic_abuse	

Table 12: Ranking of different types of implicit abuse according to perceived severity.

example 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 5 other types of implicitly abusive language from existing datasets in addition to the form of abusive language we introduced in this paper. As far as the other types of implicitly abusive language are concerned, we tried to use instances similar to those used by Wiegand and Ruppenhofer (2024). We sampled instances from exactly the same datasets as Wiegand and Ruppenhofer (2024), with one exception. Wiegand and Ruppenhofer (2024) also considered a sample of *stereotypes* from the dataset by ElSherief et al. (2021). However, since this sample mixes both negative and non-negative instances of abusive language, and within this dataset the type of stereotype is not labeled in any way, we felt using this sample would distort our experiment. Therefore, we refrained from using a sample from this dataset.

Table 11 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 12 was computed based on the proportion a particular type of implicit abuse was rated to be more severe than the other type.

Overall, the tables indicate that non-negative stereotypes were judged to be more severe than abusive euphemisms or comparisons, both of which have been examined in previous studies (Wiegand et al., 2021a, 2023). This categorization aligns closely with the findings of Wiegand and Ruppenhofer (2024), which demonstrated that norm-contravening sentences were also perceived as more abusive than both euphemisms and comparisons. Based on our new experiment, we can therefore conclude that the complete set of non-negative abusive stereotypes is rated similarly in terms of abusive severity to the subset of norm-contravening sentences.

For the sake of completeness, we describe the existing datasets from which we sampled the instances, in the following types:

Comparisons. This form of implicitly abusive language uses *like*-comparisons (35) from the dataset introduced by Wiegand et al. (2021a).

(35) You run like a headless chicken.

Euphemistic Abuse. This type of implicitly abusive utterances, such as (36), represents paraphrases of explicitly abusive utterances, such as (37), as introduced by Wiegand et al. (2023).

(36) You inspire me to fall asleep.
(37) You are boring.

Jibes. We consider the posts from Sodhi et al. (2021) extracted from the subreddit `r/roastMe` as a proxy of jibes (38).

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

Sarcasm. This form of implicit abuse is taken from ElSherief et al. (2021). All those sarcastic utterances have in common that their purpose is to attack or demean a protected class or individual (39).

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

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* (40).

(40) 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, e.g. the abusive comparisons or the instances of jibes, do not target any specific identity group.

identity group	gender	first name
Black people	female	Shanice
	male	Tyrone
gay people	female	Mary*
	male	Peter*
Jews	female	Sarah
	male	Adam
Muslims	female	Yasmin
	male	Ali
women	female	Mary
	male	N/A

Table 13: Prototypical first names for the different identity groups as used in the semantic variant *instantiation* (c.f. Table 1); *: *since there are obviously* ***no*** *stereotypical first names specifically linked to being gay, we use frequent English first names*.

E Further Details on the Crowdsourcing Experiments

Table 13 displays the full list of prototypical first names for the different identity groups as used in the semantic variant *instantiation* (c.f. Table 1).

F Stable and Unstable Characteristics

Table 14 presents a set of typical stable and unstable abusive characteristics that we identified during our experiments involving the switching of identity groups (§5). Stable characteristics are those that were consistently observed to be abusive regardless of the identity group referenced. Unstable characteristics, on the other hand, are those that, in our dataset, were only rated as abusive when associated with a specific identity group. This table complements the examples provided in §5.

G More Details on the Distribution of Abusive Language among the Different Semantic Variants

Table 15 extends Table 3 from §5 by not only displaying the percentage of sentences labeled as abusive for each variant and identity group, but also by showing distributions among potential stereotypes and prevalent characteristics. We observe that a semantic variant exerts a similar effect on the proportion of abusiveness in both subsets.

H Similarity of Identity Groups

Table 16 displays the percentage of label matches for all pairs of identity groups in our extended dataset in which identity groups were switched as detailed in Section 5.

I More Details Experimental Results

Table 17 displays all linguistic features with a precision higher than random for abusive and non-abusive (atomic) sentences. This table complements Table 5 in §4, which presents only the 8 features most strongly correlated with these 2 class labels.

Table 18 presents the performance of all classifiers in our evaluation across all semantic variants. This table complements Table 8 in §6.2, which displays only the average performance across all semantic variants and the performance on the best and worst semantic variants, specifically the *atomic* and *"some"* variants.

stable		unstable	
characteristics	**original ident. group**	**characteristics**	**original ident. group**
... have the same jobs	Muslims	... work in finances	Jews
... love clubbing	lesbians	... are basketball players	Black people
... look similar to each other	Jews	... make a good cup of tea	women
... have a fluid sexuality	women	... are sociable	gay people
... use flowery language	Muslims	... are very clever	Jews
... draw attention to themselves	gay people	... don't use swear words	women
... don't drive German cars	Jews	... are in the Navy	lesbians
... are contained	Black people	... come top of their school classes	Jews
... drive slowly	women	... smell good	women

Table 14: Examples of stable and unstable abusive characteristics. Stable characteristics remain abusive regardless of the identity group being referenced. Unstable characteristics lose their abusiveness when applied to any other identity group.

	Jews			Muslims			gay people			women			Black people			average		
		subsets			subsets			subsets			subsets			subsets			subsets	
semantic variant	**total**	**prev.**	**prev.**	**total**	**ster.**	**prev.**	**total**	**ster.**	**prev.**	**total**	**ster.**	**prev.**	**total**	**ster.**	**prev.**	**total**	**ster.**	**prev.**
atomic	48.1	61.6	12.5	22.4	31.3	1.2	63.9	85.0	27.3	65.5	82.0	25.6	36.0	40.2	31.0	47.2	60.6	19.6
implication	66.2	80.4	30.0	29.7	41.5	2.4	75.9	93.5	45.5	77.5	93.7	38.5	56.1	58.8	51.7	60.9	74.1	33.8
"always"	75.6	83.2	55.1	30.0	41.2	1.3	86.9	97.4	60.0	85.5	94.7	59.1	50.3	42.6	59.8	65.4	73.7	46.1
"all"	76.1	86.1	50.0	34.1	46.0	2.6	81.7	98.0	49.4	82.2	91.5	56.5	56.0	52.5	60.2	66.0	76.1	43.9
"many"	19.1	25.6	1.3	13.8	18.5	1.3	25.1	33.3	7.1	27.7	35.4	7.0	18.7	22.5	13.8	20.7	27.1	6.1
"some"	8.3	10.9	1.3	4.8	6.0	1.4	5.1	5.9	2.3	3.9	4.2	3.0	8.6	9.8	6.8	6.1	7.2	3.0
instantiation	6.3	8.8	0.0	7.6	10.9	0.0	10.0	11.8	6.8	5.3	5.8	4.0	12.7	12.7	12.6	8.0	9.6	4.8
authoritative report	22.3	28.9	5.0	10.5	14.4	1.2	17.7	26.1	2.4	19.9	25.9	5.1	10.1	12.7	6.9	16.5	22.4	4.1
self-identification	38.1	52.1	1.3	14.7	20.9	0.0	57.3	79.1	19.3	62.5	79.9	20.5	24.9	28.4	20.7	39.6	52.9	12.5

Table 15: Percentage of sentences labeled as **abusive**; for each semantic variant we provide this proportion among the **total** set of sentences and also among subsets (*subset of potential **ster**eotypes* / *subset of **prev**alent characteristics*).

identity group 1	**identity group 2**	**label matches in %**
Black people	Muslims	77.9
gay people	women	62.4
Jews	women	62.1
gay people	Jews	61.7
Black people	Jews	58.2
Black people	women	55.4
Jews	Muslims	55.3
Black people	gay people	53.4
gay people	Muslims	49.5
Muslims	women	48.6

Table 16: Label matches among pairs of identity groups on the extended dataset with switched identity groups (§5).

abusive (atomic) sentences (*random precision: 47.7*)		
feature	**prec**	**freq**
AUTO::GPT-4::negative_connotation (Yes)	85.7	273
MANUAL::preferences	78.1	178
AUTO::GPT-4::common (No)	75.7	411
AUTO::GPT-4::formal_language (No)	74.8	485
AUTO::GPT-4::fact (No)	66.4	740
MANUAL::work/education	63.4	82
MANUAL::competence	59.7	124
MANUAL::outward_appearance	59.3	145
AUTO::less_than_5_words	56.6	371
MANUAL::living_conditions	54.8	31
MANUAL::social_interaction	54.3	92
MANUAL::family	49.6	113
AUTO::GPT-4::positive (No)	49.1	754
not abusive (atomic) sentences (*random precision: 52.3*)		
feature	**prec**	**freq**
MANUAL::religious_practice	84.9	219
MANUAL::cultural/historical_characteristics	81.3	32
AUTO::GPT-4::fact (Yes)	79.4	501
AUTO::more_than_10_words	70.1	97
AUTO::GPT-4::formal_language (Yes)	69.3	774
AUTO::GPT-4::is_this_common (Yes)	66.6	808
MANUAL::physiology/biology/medicine	65.1	146
AUTO::GPT-4::negative_connotation (No)	62.7	985
MANUAL::food/drink	57.0	114
MANUAL::character_traits/habits	55.6	306
MANUAL::sex	55.0	20
AUTO::between_5_than_10_words	54.2	791
AUTO::GPT-4::positive (Yes)	54.2	504

Table 17: Features correlating with the 2 classes.

classifier	atomic	implic.	"always"	"all"	"many"	"some"	instant.	author.	self-id.	average
majority-class	34.3	27.5	25.3	24.9	44.2	48.4	47.9	45.5	37.5	37.3
variant_generalization	34.3	38.3	39.8	40.1	44.2	48.4	47.9	45.5	37.5	41.8
PerspectiveAPI	56.1	52.3	53.6	52.9	52.1	54.8	50.5	50.4	53.7	52.9
StereoSet*	54.2 (1.8)	53.1 (1.6)	56.8 (2.1)	53.6 (1.3)	51.7 (1.0)	50.0 (0.8)	51.9 (1.1)	54.4 (2.1)	53.3 (1.2)	53.2 (1.4)
ToxiGen	58.7	54.9	50.4	53.6	53.9	57.8	58.0	56.2	50.8	56.0
ISHate*	60.7 (2.3)	57.3 (3.0)	58.0 (2.0)	59.4 (1.7)	55.9 (0.9)	54.3 (1.4)	56.0 (1.2)	54.3 (1.6)	58.1 (2.6)	57.1 (1.9)
cross-group*	68.5 (0.5)	63.0 (1.0)	54.4 (1.0)	58.9 (1.0)	57.7 (1.0)	48.4 (0.1)	53.6 (1.8)	58.3 (1.7)	65.1 (0.7)	58.6 (1.0)
Pujari et al.*	66.2 (1.3)	63.5 (2.9)	60.6 (2.0)	65.3 (1.7)	58.7 (0.8)	55.6 (1.2)	56.7 (0.4)	57.5 (1.1)	65.0 (1.8)	61.0 (1.5)
atomic_classifier::auto*	68.3 (2.0)	67.5 (1.4)	68.4 (1.0)	67.4 (1.3)	59.2 (1.4)	51.4 (1.3)	51.9 (0.9)	59.3 (1.3)	69.5 (1.5)	62.5 (1.4)
GPT-4::zero::abusive$_{plain}$	65.9	68.7	63.1	66.4	65.0	57.5	64.2	62.9	61.1	63.9
within-group*	73.5 (0.4)	70.5 (1.3)	70.9 (1.4)	70.0 (1.3)	56.3 (1.7)	52.2 (4.3)	53.7 (6.7)	57.1 (1.1)	73.2 (0.5)	64.2 (2.1)
cross-group+within-group*	75.2 (1.1)	71.5 (1.3)	70.9 (0.9)	71.3 (1.1)	57.9 (1.5)	50.6 (4.4)	55.5 (1.7)	59.6 (1.5)	73.8 (0.8)	65.2 (1.6)
GPT-4::zero::fact	72.0	73.9	62.8	66.0	64.1	58.4	58.7	58.7	73.5	65.4
GPT-4::zero::formal	71.3	73.8	69.3	70.5	63.0	62.1	59.2	58.4	66.2	66.0
GPT-4::zero::neg._connot.	70.2	68.6	68.1	68.0	64.7	60.4	**64.8**	**64.6**	64.8	66.0
GPT-4::zero::common	68.9	74.4	66.8	68.8	62.7	56.7	64.2	61.4	63.9	65.3
GPT-4::zero::abusive$_{sensitive}$	72.7	74.6	69.5	69.9	**66.4**	**63.4**	63.6	62.6	64.1	67.4
GPT-4::augmenting*†	**78.3** (0.9)	**75.9** (1.2)	**75.3** (0.8)	**75.3** (1.1)	64.5 (0.9)	62.0 (2.0)	60.0 (1.6)	62.7 (0.8)	**77.7** (0.2)	**70.2** (1.0)
atomic_classifier::oracle	100.0	77.0	72.3	75.2	69.2	59.6	60.4	66.5	77.5	73.1
human_classifier	82.1 (0.5)	80.1 (0.7)	80.9 (1.3)	80.1 (1.4)	75.1 (0.9)	69.8 (2.3)	71.6 (0.8)	72.6 (0.4)	79.7 (0.8)	76.9 (1.0)

Table 18: F1-score of classifiers on each semantic variant; numbers in brackets denote standard deviation; *: fine-tuned with DeBERTa; †: uses the completions of the 5 best zero-shot approaches (*prompts are described in Tables 4 & 7*).