

# Beyond the Explicit: A Bilingual Dataset for Dehumanization Detection in Social Media

Dennis Assenmacher<sup>1\*</sup>, Paloma Piot<sup>2</sup>,  
Katarina Laken<sup>3</sup>, David Jurgens<sup>4</sup>, Claudia Wagner<sup>1,5</sup>

<sup>1</sup>GESIS – Leibniz Institute for the Social Sciences,

<sup>2</sup>IRLab, CITIC Research Centre, Universidade da Coruña,

<sup>3</sup>Fondazione Bruno Kessler & Universidade de Santiago de Compostela,

<sup>4</sup>University of Michigan, <sup>5</sup>RWTH Aachen University

dennis.assenmacher@gesis.org

## Abstract

**Content Warning:** This paper contains examples of dehumanizing and offensive language.

Digital dehumanization, although a critical issue, remains largely overlooked within the field of computational linguistics and Natural Language Processing (NLP). To our best knowledge, this work presents the most comprehensive annotated bi-lingual dataset on dehumanization, with almost 16.000 instances (44% dehumanizing) that exists today and can serve as both a training resource for machine learning models and a benchmark for evaluating future dehumanization detection techniques.

Our comparison of 15 methods shows that traditional hate speech detection approaches perform poorly on dehumanization detection, particularly for implicit forms. By leveraging our dataset in few-shot and fine-tuning settings, LLMs achieve comparable performance on implicit and explicit dehumanization and substantially outperform zero-shot baselines, with average gains of 39% and 65%, respectively. However, detection accuracy remains slightly lower for mechanistic than for animalistic dehumanization, the latter being the dominant focus of prior work.

## 1 Introduction

Dehumanization is a psychological process in which individuals or groups are described or treated as less than fully human, often by likening them to animals, objects, or machines. This process de-emphasizes the humanity of the targeted group either partially or entirely, thereby weakening moral obligations toward them (Haslam, 2006). As a result, dehumanization can lead to extreme inter-group bias, heightened anxiety, hate speech, and violence directed at specific social groups. It often facilitates moral exclusion, where certain populations are perceived as existing outside the

boundaries of moral concern. Such perceptions have historically enabled atrocities, war crimes, and genocides. Given these severe consequences, the ability to accurately measure different forms of dehumanization becomes crucial. For instance, systematic detection can support moderators of online platforms in raising user awareness and/or removing harmful content. Beyond moderation, monitoring dehumanizing language at scale can also help identify rising intergroup tensions or spikes in harmful rhetoric, serving as early warning signals for potential real-world conflicts or unrest.

We distinguish between dehumanization, hate speech, and offensive language, which are partly overlapping but not equivalent constructs. We define dehumanization as language that represents a target as less than fully human, either through explicit comparisons to non-human entities (e.g., animals, objects) or by implicitly denying human traits such as rationality or morality, following the dual model of dehumanization (Haslam, 2006). There is no unique agreed-upon definition for hate speech but hate speech definitions often refer to language that attacks, derogates, humiliates or discriminates a target based on group membership and protected characteristics such as gender, ethnicity, sexual orientation (Melis et al., 2025). For offensive content we adapt the definition of Zampieri et al. (2019) that define offensive posts as insults, threats, and messages containing any form of untargeted profanity. Examples of these constructs are provided in Appendix Table A.

In the field of computational linguistics and NLP only recently researchers started developing methods to detect dehumanization and mainly focused on animalistic dehumanization and explicit forms of dehumanization. Recognizing this gap, we present the largest dehumanization dataset ( $\approx$  16.000 instances, 44% dehumanizing) that covers two different languages, two different social media

\*Corresponding author.

platforms and the two theoretical dimensions of dehumanization defined by the Dual Dehumanization Theory (Haslam, 2006). Using three different sampling methods, we investigate the coverage of the two dimensions in its explicit and implicit form.

Our comparison of 15 methods shows that traditional hate speech detection approaches perform poorly on dehumanization detection, particularly for implicit forms. This highlights the need for research on detecting dehumanization despite it partly overlaps with hatespeech. In contrast, LLMs achieve comparable performance on both implicit and explicit dehumanization when applied in few-shot or fine-tuning settings with significant performance gains over corresponding zero-shot scenarios. Our best-performing model, a fine-tuned Llama 3.1 70B model, achieves an average  $F_1$  score of 0.77 across all dehumanization dimensions and languages. However, LLMs still show slightly lower performance in detecting mechanistic dehumanization than animalistic dehumanization.

## 2 Background and Related Work

**Theories of Dehumanization** Dehumanization has been extensively studied in psychology, leading to the development of various theories. A thorough report of existing dehumanization theories can be found in Haslam and Loughnan (2014). Our work builds upon the Dual Model of Dehumanization, proposed by Haslam (2006), that distinguishes between two forms: animalistic and mechanistic dehumanization. Additionally, Haslam differs between subtle and explicit forms of dehumanization (Haslam, 2013). Explicit animalistic dehumanization is characterized by direct comparisons with non-human living creatures such as animals or vermin. Implicit animalistic dehumanization involves the denial of attributes that are unique to humans, such as intelligence, civility, rationality, or morality. In contrast, implicit mechanistic dehumanization involves the denial of attributes that are inherent to human nature, including emotional responsiveness, warmth, and agency. Explicit mechanistic dehumanization involves judgments that refer directly to inanimate objects, like robots, machines or scum. Figure 1 depicts the four dimensions of dehumanization that we derive from Haslam’s work.

It should be emphasized that not all explicit and direct comparisons to animals or machines are nec-

essarily dehumanizing in a sense that they cause harm to people (Over, 2021). *He is strong like a lion* or referring to an athlete as a *machine* are direct animalistic/mechanistic comparisons that have a positive connotation.

**Measuring Dehumanization** Prior computational work on dehumanization largely falls into three categories: (1) lexical approaches relying on dictionaries or keyword signals (Mendelsohn et al., 2020; Landry et al., 2022; Schweitzer and Waytz, 2021); (2) contextual approaches that infer dehumanization from masked-token predictions or contextual embeddings (Card et al., 2022; Engelmann et al., 2024); and (3) LLM or concept driven approaches that identify metaphorical or implicit dehumanizing content with minimal manual annotation (Saffari et al., 2025; Friedman et al., 2021; Mendelsohn and Budak, 2025). A notable exception is the recent work by Saffari et al. (2025) that shows that LLMs (especially Claude) can achieve high performance on dehumanization detection - the dataset that they use focuses however on explicit animalistic dehumanization which is the easiest detectable dimension of dehumanization.

Our work extends this literature by adapting state of the art language models to detect different theoretical dimensions of dehumanization in two languages, across two platforms, and for a broader range of target groups than prior research.

**Datasets** Existing datasets on dehumanization differ strongly in size, annotation schemes, theoretical focus, and language coverage (Burovova and Romanyszyn, 2024; Engelmann et al., 2024; Sachdeva et al., 2022; Vidgen et al., 2021; Kirk et al., 2023; Ho, 2022; Vidgen et al., 2020; Joshi, 2025; Saffari et al., 2025). Many focus on English, rely on binary or highly explicit labels, and center on the animalistic dimension of dehumanization.

Our dataset captures multiple theoretical dimensions of dehumanization beyond purely animalistic forms, includes both document- and span-level annotations, spans two platforms, two languages, and covers more than a decade.

## 3 The Dataset(s)

### 3.1 Data Sources

The datasets utilized in this study originated from a comprehensive  $\mathbb{X}$  dataset, formerly known as Twitter, collected over a span of 10 years (from

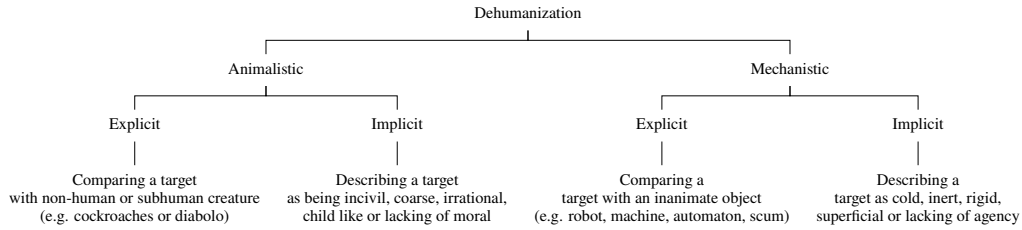


Figure 1: Conceptualization of Dehumanization based on (Haslam, 2006, 2013)

2013 to 2023) (Fafalios et al., 2018). This dataset comprises 1% of all tweets, encompassing data from January 2013 onwards, obtained through the public Twitter API. Furthermore, this Twitter data was enriched by four years (2019-2022) of Decahose content (10% data stream). We also draw on Reddit data collected via the Pushshift API over the same period (2013–2023) (Baumgartner et al., 2020). The Pushshift dataset includes Reddit posts, comments, and associated metadata; however, for this study, we focus exclusively on user-generated comments. We restrict our study to tweets and Reddit comments posted in English or German. Also we only consider original content, i.e. no quotes or retweets.

### 3.2 Candidate Selection

Our study investigates a broad spectrum of potentially dehumanizing messages. To manage the large volume of data, we initiated the analysis by filtering messages that reference possible targets of dehumanization. This included explicit mentions of specific groups, as well as messages using personal pronouns like “they” or demonstrative pronouns like “those” and “these” to imply group references.

The list of target groups is based on prior research identifying populations particularly vulnerable to dehumanization (Hodson and Costello, 2007; Goff et al., 2008; Vaes et al., 2011; Loughnan et al., 2014), and was expanded through manual inspection and semantic similarity using fast-Text embeddings trained on our dataset. This allowed us to capture direct mentions, variations, and colloquial forms. In total we used 184 (156 German) target terms for 9 target groups. The final lists of target terms for both languages are provided in Appendix Tables 3,4.

### 3.3 Dehumanization Candidates Selection

After filtering candidate messages from Twitter and Reddit that potentially refer to targets, we ap-

ply three different sampling methods to select instances for human annotations. The goal of our sampling strategies is to increase the prevalence of different forms of dehumanization in our dataset that humans annotate. We use *theory-informed instance selection* that relies on survey scales for dehumanization, *dataset-informed instance selection* which relies on previous dehumanization datasets and *LLM-informed instance selection* which relies on LLM-generated examples of dehumanization. All selection methods use SBERT to compute semantic similarities between the potential candidates and some seeds (e.g., survey items). The selection methods and their ability to find relevant candidates are described in the Appendix E.

**Selecting Annotation Candidates** Using our three different instance selection approaches, we gathered 349,916 potential Twitter and 122,641 potential Reddit candidates for crowd annotation. From these candidates, we sampled potentially explicit and implicit instances per language for annotation (see Table 1), with sampling probabilities weighted by the average of the corresponding SBERT similarities calculated in the selection steps.

### 3.4 Annotation Procedure

We collected span-level and document-level annotations<sup>1</sup> for the annotation subset using Prolific, a crowd-working platform, based on the taxonomy shown in Figure 1. Before launching the main annotation study, we conducted several pilot studies (in our lab as well as on Prolific) that tested different task designs, instructions, and user interfaces. These pre-studies helped us assessing task clarity and annotator agreement. Based on these outcomes, we selected the most effective setup for the main annotation study.

Each candidate instance was annotated by at least three independent crowd-workers. To ensure annotation quality, we required annotators to be flu-

<sup>1</sup>The task interface can be seen in Appendix C.1

ent in English (or German), familiar with social media platforms, and hold at least a high school degree or equivalent. For the English task, we restricted participation to crowdworkers with an approval rate of at least 98% to ensure high annotation quality. For the German task, this criterion could not be applied, as the available pool of native-speaking annotators was considerably smaller and would have rendered recruitment infeasible<sup>2</sup>. Annotators were asked to perform the following three tasks<sup>3</sup>:

1. Determine whether the post was perceived as an offensive statement directed at an individual or group.
2. Highlight any portion of the text that exhibits animalistic or mechanistic dehumanization (see Appendix C for definitions and examples).
3. Identify and highlight the target of the dehumanizing expression.

We conducted 8 separate annotation studies, one for each dimension (explicit/implicit animalistic, explicit/implicit mechanistic) and language (German/English). Each annotation task was setup as a separate annotation task since annotation instructions were very detailed for each dimension (see Appendix C). For example, within the explicit dehumanization task annotators had to decide if a message compares a target group to an animal, a disease, a subhuman creature, or an inanimate object. If nothing applied the message is considered as “not explicit animalistic dehumanization”. If an animal, disease or subhuman creature comparison is highlighted the message is considered “explicit animalistic dehumanization”. If a comparison with an inanimate object is found, the instance is an example of “explicit mechanistic dehumanization”. In theory one instance can contain multiple forms of dehumanization (e.g., animalistic and mechanistic dehumanization), but this did not happen in practice. For quality assurance, we embedded attention-check questions and excluded all crowd-worker annotations from our dataset when workers consistently failed these checks.

After annotation, individual judgments were aggregated using majority voting. Importantly,

agreement on the top-level category did not require agreement on the specific subcategory. Thus, annotators could agree that an instance belongs to a broader dehumanization category while differing in their interpretation of the more fine-grained subtype. For example, an instance could be labeled as explicit animalistic dehumanization even if individual annotators respectively categorized it as an animal comparison, subhuman comparison, or disease comparison.

**Annotation Results** Table 1 summarizes the annotation results and the proportion of instances labeled as dehumanizing (majority vote) across languages and dimensions. Overall, our selection strategies yielded a substantial share of dehumanizing instances in both English and German. However, it is noteworthy that explicit mechanistic dehumanization is comparatively rare in both languages (English: 9.8%, German: 5.5%), which may indicate a lower prevalence of mechanistic dehumanization on  $\mathbb{X}$  and Reddit compared to animalistic dehumanization.

Although a substantial proportion of candidate instances are labeled as offensive across all dimensions, dehumanization does not coincide with offensiveness. This is especially pronounced in the implicit dimension, where 15% of positively labeled animalistic and 33% of mechanistic instances are dehumanizing while not being perceived offensive by the majority of annotators.

**Rater Agreement** We report Inter-Annotator Agreement (IAA) using Krippendorff’s  $\alpha$  (cf. Table I). For explicit dehumanization, the English dataset exhibits an IAA of 0.393, while the German dataset shows a higher level of agreement ( $\alpha = 0.475$ ). Examining theoretical sub-dimensions reveals that IAA for explicit animalistic dehumanization in English is slightly higher than for mechanistic dehumanization (0.431 vs. 0.381). Within the animalistic dimension, the animal comparison category shows comparatively strong agreement (IAA = 0.577), whereas subhuman, one of the conceptually more complex categories, exhibits the lowest agreement (IAA = 0.275).

For implicit dehumanization, Krippendorff’s  $\alpha$  is 0.290 (English) and 0.517 (German) for animalistic dehumanization, and 0.292 (English) and 0.289 (German) for mechanistic dehumanization. These relatively low scores have been observed in previous studies on dehumanization (Sachdeva et al., 2022) and can be attributed to several fac-

<sup>2</sup>Annotator statistics in Appendix H

<sup>3</sup>Schematic diagram of the annotation process can be found in Appendix 5

| Dehumanization   | English |        |        |        |       |        | German  |        |        |        |       |        |
|------------------|---------|--------|--------|--------|-------|--------|---------|--------|--------|--------|-------|--------|
|                  | Twitter |        | Reddit |        | Total |        | Twitter |        | Reddit |        | Total |        |
|                  | #       | %      | #      | %      | #     | %      | #       | %      | #      | %      | #     | %      |
| Total            | 2215    |        | 1772   |        | 3987  |        | 3374    |        | 477    |        | 3851  |        |
| Not Dehumanizing | 1257    | 56.60% | 1152   | 64.76% | 2409  | 60.22% | 2370    | 67.3%  | 351    | 73.60% | 2721  | 67.30% |
| Explicit         | 964     | 43.40% | 627    | 35.24% | 1591  | 39.77% | 1150    | 32.70% | 126    | 26.40% | 1276  | 32.70% |
| Animalistic      | 658     | 29.70% | 426    | 24.04% | 1084  | 26.56% | 819     | 24.27% | 100    | 21.00% | 919   | 22.60% |
| Mechanistic      | 240     | 10.10% | 167    | 9.42%  | 508   | 9.80%  | 193     | 6.00%  | 24     | 5.00%  | 217   | 5.50%  |
| Offensive        | 1292    | 58.32% | 847    | 47.80% | 2139  | 53.64% | 1616    | 47.90% | 186    | 39.00% | 1802  | 46.80% |
| Total            | 1558    |        | 2442   |        | 4000  |        | 3458    |        | 542    |        | 4000  |        |
| Not Dehumanizing | 549     | 35.24% | 1089   | 44.59% | 1638  | 40.95% | 1773    | 51.27% | 364    | 67.16% | 2137  | 53.42% |
| Implicit         | 1009    | 64.76% | 1353   | 55.41% | 2362  | 59.05% | 1685    | 48.73% | 178    | 32.84% | 1863  | 46.58% |
| Animalistic      | 640     | 41.78% | 731    | 29.93% | 1371  | 34.27% | 1112    | 32.16% | 67     | 12.36% | 1179  | 29.47% |
| Mechanistic      | 369     | 23.68% | 622    | 25.47% | 991   | 24.77% | 573     | 16.57% | 111    | 20.48% | 684   | 17.10% |
| Offensive        | 886     | 56.87% | 1123   | 45.99% | 2009  | 50.22% | 1794    | 51.88% | 181    | 33.39% | 1975  | 49.37% |

Table 1: Annotated Dataset description. The explicit and implicit dataset consist each of approx. 4.000 unique instances, both categorized through span-annotations and majority voting. Additionally, Offensive annotations were made independently on a document-level.

tors. First, the perception of dehumanization is inherently subjective. Individual annotators may vary in their sensitivity to particular expressions, depending on their background, values, or lived experiences. Second, without sufficient context, it is often difficult to determine whether a statement is intended as dehumanizing. For example, the statement “*Not, but see because they’re actually maggots over there*” could either imply a comparison between people and maggots or simply describe the presence of maggots in a certain location, making its interpretation ambiguous. We emphasize that this annotation task is inherently challenging, particularly because all instances were selected as already being potential candidates for dehumanization. Moreover, Wong et al. (2021) argue that in tasks involving subjective judgment, like identifying dehumanizing language, a Cohen’s kappa threshold of 0.6 is too strict, since annotators often vary in cultural background and training.

**Expert Annotations** For a stratified random sample comprising 20% of the dataset (used as test data in our experiments), three expert annotators (the authors of this paper who hold PhDs and master degrees in Computer Science, Computational Linguistics and Computational Social Science) labeled the instances using the same guidelines provided to crowdworkers. Their annotations yielded an agreement level comparable to that of the crowdworkers. However, in contrast to the crowdworkers, the experts resolved remaining disagreements through several rounds of discussion.

## 4 Methods and Experimental Setup

We define our task as a binary dehumanization detection task, train for each language various models on each of the four dimensions of dehumanization and compare them with out-of-the-box models that do not require any supervision. Here, we rely on methods that are frequently used for hate speech detection such as fine-tuned BERT (bert-base-uncased for English and bert-base-multilingual-uncased for German) (Devlin et al., 2019) and RoBERTa (xml-roberta-large) (Conneau et al., 2020) models, alongside established hate speech detection architectures HateBERT (Caselli et al., 2021) and MetaHateBERT (Piot et al., 2024)<sup>4</sup>. Among the established hate speech detection models, both were fine-tuned on our dataset for 3 epochs with a learning rate of 5e-5, batch size of 32, weight decay of 0.18, and Adam 8bit optimizer.

In addition to our experiments with BERT-based model architectures, we also tested generative large language models (LLMs) based on a decoder-only architecture. We examined Llama 3.1 8B (Llama-3.1-8B-Instruct) and Llama 3.1 70B (Llama-3.1-70B-Instruct) (Grattafiori et al., 2024) in zero-shot and few-shot (20 shots) settings with the same instructions given to the human annotators (see Appendix C). Furthermore, we fine-tuned the LLMs via LoRA (Hu et al., 2021) for 4 epochs using a learning rate of 2e-4,

<sup>4</sup>As dehumanization is a subconstruct of hate speech we expect that these models are able to capture some of its signals

batch size of 4, weight decay of 0.01 and AdamW torch optimizer. Evaluation was performed using greedy decoding (temperature = 0, no sampling), such that the model always produced the most likely continuation. Finally, we use black-box models, the Perspective API (toxicity score) and GPT 4.1, that we do not fine-tune or train on our data (Lees et al., 2022). We also test the performance of MetaHateBERT and HateBERT without fine-tuning on our data (vanilla setup)<sup>5</sup>. We compare the performance of methods on fixed hold-out splits for each language and dehumanization dimension (20%). Experiments were repeated five times with different random seeds. All experiments were conducted on a single NVIDIA A100 80GB instance requiring 100 GPU/h.

## 5 Results

The performance of different models and model architectures for each of the four dimensions of dehumanization is shown in Figure 2 for English (and in Appendix Figure 9 for German). In the following we discuss general findings that apply to both languages and highlight differences between them.

**Out-of-the-Box Performance** Models that have not been exposed to any data from our dataset provide a useful baseline for evaluation. These include both models pretrained on related hate speech or abusive language tasks—such as Perspective API, HateBERT (vanilla), and MetaHateBERT (vanilla)—and general-purpose language models like Llama 3.1 (zero-shot) and GPT-4.1.

Baseline toxicity classifiers, such as the Perspective API, HateBERT, and MetaHateBERT, perform reasonably well on instances of animalistic dehumanization with strong negative valence. However, these models tend to fail when the dehumanization is more implicit and lacks explicitly negative wording (see Figure 12) like *“My dogs are Republicans too, when one barks for no reason they all bark.”*. Such indirect and figurative expressions often go undetected by these standard toxicity models.

When LLMs are used in a zero-shot setting, the performance is limited. In the zero-shot setting, GPT-4.1 outperforms the open-source Llama 3.1 variants but still shows limited overall performance. Its only notable success appears in the implicit mechanistic category, where it achieves a relatively high  $F_1$  score of 0.72.

<sup>5</sup>Both models only work on English data. Thus we do not test them on our German subset

**Few-Shot and Fine-Tuned Models** Models trained on our dataset achieve substantial performance improvements across all dehumanization categories. On average, LLMs show a 65% relative  $F_1$  increase across categories and languages when moving from zero-shot ( $F_1 = 0.5$ ) to fine-tuned ( $F_1 = 0.75$ ) models. Even in the few-shot setting ( $F_1 = 0.65$ ), where only 20 examples are provided in the prompt, performance improves by an average of 39% compared to the corresponding zero-shot condition. Further fine-tuning of existing BERT-based hate speech classifiers, such as MetaHateBERT, on our dataset also leads to notable gains in all dehumanization dimensions. For example, MetaHateBERT’s average  $F_1$  score increases by 19% when the model is fine-tuned ( $F_1 = 0.69$ ) on our data compared to the vanilla model ( $F_1 = 0.57$ ) that was trained on 39 existing hate speech-related datasets. The best average results across all four dimensions were achieved by the fine-tuned Llama 3.1-Instruct models (Llama 3.1 70B fine-tuned  $F_1 = 0.77$  and Llama 3.1 8B fine-tuned  $F_1 = 0.76$ , followed by fine-tuned HateBERT  $F_1 = 0.74$  and GPT-4.1 few-shot  $F_1 = 0.73$ ). These models all perform better on animalistic dehumanization compared to mechanistic dehumanization (cf. Figure 10) suggesting that mechanistic dehumanization is more difficult to detect than animalistic dehumanization. One potential explanation is the lower prevalence of mechanistic dehumanization (cf. Table 1). Overall, fine-tuning on our data proved beneficial across all dimensions, with the largest gains observed for the implicit and explicit animalistic categories. The former benefits from learning to recognize indirect animal comparisons, while the latter improves due to the model’s exposure to more diverse examples of denials of unique human traits.

While the larger 70B Llama model does on average outperform the smaller 8B variant, the performance difference is relatively modest in the fine-tuning setting and for some dimensions (e.g. explicit animalistic) the 8B fine-tuned model even outperforms its larger counterpart. This result is particularly relevant for practical applications, as the 8B model is significantly more cost-efficient, environmentally friendly, and easier to deploy. It is therefore a viable option for scenarios with limited computational resources (Piot and Parapar, 2025). However, in the zero-shot setting, the larger model variants exhibit a clear advantage, suggest-

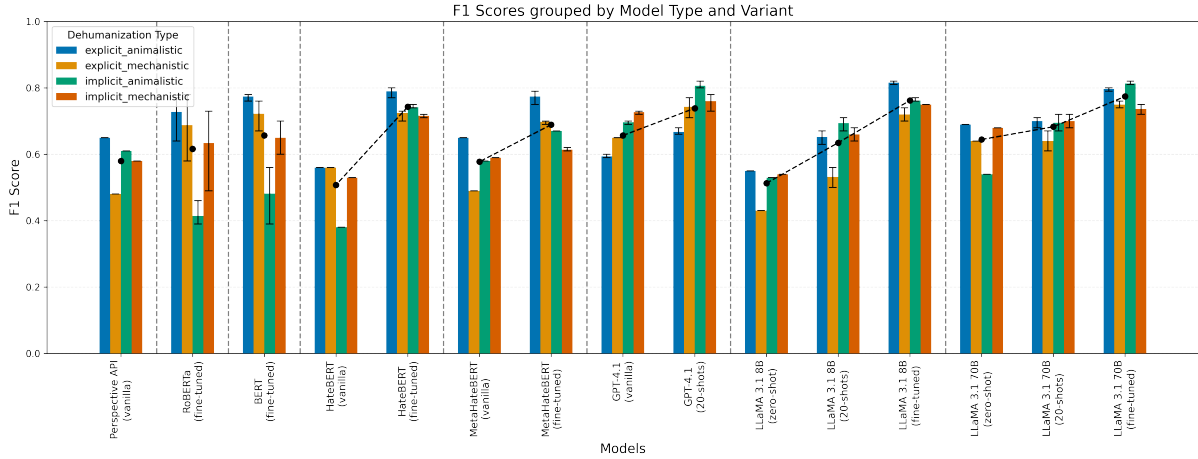


Figure 2: Test  $F_1$ -Scores for the four different dimensions of dehumanization (see color bars) on English Reddit and Twitter data for different model architectures.

ing that they possess a stronger inherent understanding of the underlying constructs and therefore achieve higher performance without task-specific supervision.

**Reproducibility** Code, target lists, templates and artificially generated instances are publicly available on GitHub.<sup>6</sup> Document-level and pseudonymized span-level annotations of real posts, as well as the corresponding message IDs, are available upon registering with the GESIS Data Archive.

**Error Analysis** For our error analysis, we focus on instances that remain misclassified even when all top-performing systems agree. We select the top-performing models based on their average test-set F1 score across all runs and dehumanization dimensions. We include fine-tuned MetaHateBERT, GPT-4.1 with 20-shot prompting, and fine-tuned Llama-3.1-8B and Llama3.1-70B-Instruct. These four models consistently rank among the best performers across the four dimensions of the construct. For each model, we aggregate multiple runs and compare the majority vote prediction of the model with the gold (expert) label. We retain only instances where all four models are wrong (unanimous errors); this yields a conservative set of 127 cases (37 explicit animalistic, 26 explicit mechanistic, 22 implicit animalistic, 35 implicit mechanistic). Figure 3 shows that false negatives are more common for explicit dehumanization (especially animalistic: 35 FN vs. 2 FP), while models for the implicit dimensions reveal more false

|                      |    |    |
|----------------------|----|----|
| Explicit animalistic | 2  | 35 |
| Explicit mechanistic | 0  | 26 |
| Implicit animalistic | 18 | 4  |
| Implicit mechanistic | 32 | 3  |
|                      | FP | FN |

Figure 3: Error Types identified across consistently misclassified instances.

positive errors (e.g., implicit animalistic: 18 FP, 4 FN).

We observe a relatively high number of false negatives for the explicit dimensions and a relatively high number of false positives for the implicit dimensions. A qualitative inspection of the error cases in the explicit dehumanization dimensions, suggests that the model learned a comparatively focused and lexicalized representation of this phenomenon. While human annotators frequently labeled broader and sometimes implicit forms of degradation, disgust framing, pathology-based derogation, or subhuman comparisons as explicitly dehumanizing, the model appeared to rely primarily on explicit non-human or animal/disease-related cues such as “parasites,” or “less than human.” As a result, many hostile or degrading statements that lacked direct animalistic terminology were classified as non-dehumanizing, despite being annotated as positive examples. We further observe that highly sexualized statements targeting women - examples of mechanistic dehumanization - were often not correctly identified, despite

<sup>6</sup><https://github.com/Dennis1989/Beyond-the-Explicit-Dataset>

portraying individuals in degrading and potentially subhumanizing ways.

For implicit dehumanization, we observe that the models appear to apply a broader interpretation than the expert annotators, frequently associating general derogation or implied inferiority with dehumanization. For example, the statement “*I mean come on the dude doesn’t even know anything about space even.*” was classified by the model as implicitly dehumanizing due to its portrayal of the target as unintelligent or incapable, whereas expert annotators labeled the statement as non-dehumanizing. Similar patterns can be observed for statements such as “*Those people are probably functionally illiterate*” or “*These silly rednecks*” where the models appeared to associate degrading assumptions about intelligence, civility, or rationality with implicit subhumanization more readily than the annotators. In these cases, annotators may not have considered the threshold for fundamentally denying intelligence, rationality, or human competence to be sufficiently reached, despite the presence of derogatory or stereotypical framing.

These findings illustrate that the interpretation of dehumanizing is sometimes subjective and is sensitive to contextual interpretation. For example, statements such as “*Asylum seekers in my country are a fucking nuisance*” may be interpreted by some annotators as implicitly portraying a social group as burdensome, socially undesirable, or less worthy, while others may regard the statement as offensive or hostile rather than explicitly dehumanizing. This interpretation is further reflected by the fact that 52% of the misclassified instances stem from cases in which expert annotators initially disagreed before disagreement resolution was performed.

## 6 Discussion

The review of existing approaches to measuring dehumanizing content revealed several limitations, most notably the scarcity of datasets, their narrow focus on explicit, often animalistic comparisons, and their predominant focus on the English language. To address this gap, we annotated a large, bilingual dataset from Twitter and Reddit that covers the different theoretical dimensions of dehumanization and stretches over more than a decade.

Our dataset includes both dehumanizing and non-dehumanizing examples, often topically or lin-

guistically similar due to a similarity-based retrieval strategy. Consequently, even negatively labeled examples are plausible dehumanization candidates, making the dataset challenging and valuable for both training and rigorous evaluation. An additional advantage of our dataset is the inclusion of span-level annotations, allowing for a more precise identification of dehumanizing language within text that can be used in the future.

Our baseline experiments show that current general-purpose LLMs, as well as classifiers pre-trained on hate speech tasks, struggle to identify the different dimensions of dehumanization and benefit from utilizing our dataset in few-shot or fine-tuning scenarios. We compare the performance of 15 different models on the binary dehumanization detection for each sub-dimension of the construct. Our results reveal that most methods (except GPT-4.1) achieve their best performance on explicit animalistic dehumanization in the English data. In the German data we see that explicit and implicit animalistic dehumanization are the best performing dimensions for all models including GPT-4.1. In general, vanilla and zero-shot models perform substantially worse than fine-tuned or few-shot models. The only exception here is GPT-4.1, which also shows good performance in the zero-shot setting. We find that all models trained on our dataset show substantial performance gains across all dehumanization categories.

Beyond these empirical findings, our results suggest a broader conceptual implication: dehumanization cannot be fully captured as a purely surface-level linguistic phenomenon. The strong performance of models on animalistic forms, contrasted with their difficulty in identifying mechanistic and/or more implicit forms, indicates that dehumanizing content often relies on deeper semantic and normative cues. This pattern is consistent with the view that dehumanization reflects underlying moral processes, such as the exclusion of certain groups from moral concern. It also suggests that future modeling efforts could benefit from incorporating representations of moral reasoning, such as the denial of agency, dignity, or rights, to more accurately capture the full spectrum of dehumanizing content.

## 7 Limitations

We aimed to create a comprehensive dataset by incorporating a wide range of dimensions and fre-

quently targeted groups that goes beyond specific target groups that are addressed in current research. However, we cannot ensure that all possible targets of dehumanization are included and our dataset is limited to the list of target identifiers that we used. To address this limitation to some extent, we complemented the target list with personal pronouns like “they” and demonstrative pronouns like “those” and “these” that imply group references.

Our dataset is also limited to two platforms (Twitter (X) and Reddit) and two languages (German and English) and is not a random or representative sample of the content that was published at these platforms between 2013 and 2023. However our dataset provides a diverse set of examples of dehumanizing messages that were published over the course of ten years. A further limitation concerns the span annotations included in the dataset. While annotators highlighted spans associated with dehumanizing expressions, the current work does not evaluate span-level inter-annotator agreement or model performance on the extraction task itself. As a result, the reliability and consistency of the span annotations remain an open question and should be investigated more systematically in future work.

Modeling the explicit mechanistic dimension, especially in German, was challenging due to a lack of training instances, likely reflecting the lower prevalence of mechanistic dehumanization compared to animalistic forms. Future efforts should focus on gathering more examples in this category.

Furthermore, dehumanization often occurs not only through text, but also through images or memes. Since our dataset includes only textual content, we capture text-based signals only.

## Acknowledgements

This work was created in the context of the project: Digital Dehumanization: Measurements, Exposure and Prevalence (DeHum), funded by the Leibniz Association Competition. The authors thank the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. The second author also thanks the financial support supplied by the grants PID2022-137061OB-C21 and PID2025-167749OB-C22 funded by MI-CIU/AEI/10.13039/501100011033 and by “ERDF/EU”. The second author also thanks the funding supplied by the Consellería de

Cultura, Educación, Formación Profesional e Universidades (accreditations ED431G 2023/01 and ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

## References

- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.
- Kateryna Burovova and Mariana Romanyshyn. 2024. [Computational analysis of dehumanization of ukrainians on Russian social media](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 28–39, St. Julians, Malta. Association for Computational Linguistics.
- Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustán, Ran Abramitzky, and Dan Jurafsky. 2022. [Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration](#). *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. [HateBERT: Retraining BERT for abusive language detection in English](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Paul Engelmann, Peter Trolle, and Christian Hardmeier. 2024. [A dataset for the detection of dehumanizing language](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 14–20, St. Julian’s, Malta. Association for Computational Linguistics.
- Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. [Tweetskb: A public and large-scale rdf corpus of annotated tweets](#). In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*, page 177–190, Berlin, Heidelberg. Springer-Verlag.
- Scott E Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Patel Pooja, and Christopher Miller. 2021. Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.
- Phillip Atiba Goff, Jennifer L Eberhardt, Melissa J Williams, and Matthew Christian Jackson. 2008. [Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences](#). *Journal of personality and social psychology*, 94(2):292–306.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Nick Haslam. 2006. [Dehumanization: An integrative review](#). *Personality and Social Psychology Review*, 10(3):252–264. PMID: 16859440.
- Nick Haslam. 2013. [What is dehumanization?](#) In Paul G. Bain, Jeroen Vaes, and Jacques-Philippe Leyens, editors, *Humanness and Dehumanization*, pages 34–48. Psychology Press.
- Nick Haslam, Stephen Loughnan, Yoshihisa Kashima, and Paul Bain. 2008. Attributing and denying humanness to others. *European review of social psychology*, 19(1):55–85.
- Nick Haslam and Steve Loughnan. 2014. [Dehumanization and Infrahumanization](#). *Annual Review of Psychology*, 65(1):399–423.
- Janet Ho. 2022. [Evading the lockdown: Animal metaphors and dehumanization in virtual space](#). *Metaphor and Symbol*, 37(1):21–38.
- Gordon Hodson and Kimberly Costello. 2007. [Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes](#). *Psychological science*, 18(8):691–698.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
- Devavrat Joshi. 2025. [Decoding dehumanization: Leveraging nlp to identify dehumanizing language and its targets](#). In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval, NLPPIR ’24*, page 87–96, New York, NY, USA. Association for Computing Machinery.
- Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. [SemEval-2023 task 10: Explainable detection of online sexism](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.
- Olivia R. Kuljian and Zachary P. Hohman. 2023. [Warmth, competence, and subtle dehumanization: Comparing clustering patterns of warmth and competence with animalistic and mechanistic dehumanization](#). *British Journal of Social Psychology*, 62(1):181–196.
- A. P. Landry, R. I. Orr, and K. Mere. 2022. [Dehumanization and mass violence: A study of mental state language in nazi propaganda \(1927-1945\)](#). *PLoS ONE*, 17(11):e0274957.
- Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. [A new generation of perspective api: Efficient multilingual character-level transformers](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’22*, page 3197–3207, New York, NY, USA. Association for Computing Machinery.
- Steve Loughnan, Nick Haslam, {Robbie M.} Sutton, and Bettina Spencer. 2014. [Dehumanization and social class animality in the stereotypes of “white trash,” “chavs,” and “bogans”](#). *Social Psychology*, 45(1):54–61.
- Matteo Melis, Gabriella Lapesa, and Dennis Assenmacher. 2025. [A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance](#). In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 490–521, Vienna, Austria. Association for Computational Linguistics.
- Julia Mendelsohn and Ceren Budak. 2025. [When people are floods: Analyzing dehumanizing metaphors in immigration discourse with large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*

- 1: *Long Papers*), pages 8079–8103, Vienna, Austria. Association for Computational Linguistics.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. [A framework for the computational linguistic analysis of dehumanization](#). *Frontiers in Artificial Intelligence*, 3.
- Harriet Over. 2021. [Seven Challenges for the Dehumanization Hypothesis](#). *Perspectives on Psychological Science*, 16(1):3–13.
- Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. [Metahate: A dataset for unifying efforts on hate speech detection](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):2025–2039.
- Paloma Piot and Javier Parapar. 2025. [Towards Efficient and Explainable Hate Speech Detection via Model Distillation](#), page 376–392. Springer Nature Switzerland.
- Pauline Rasset, Benoît Montalan, and Jessica Mange. 2023. [Only human after all? a pre-registered study on gaze behavior and humanity attributions to people with facial difference](#). *PLOS ONE*, 18(12):1–8.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Hamidreza Saffari, Mohammadamin Shafiei, Hezhao Zhang, Lasana T. Harris, and Nafise Sadat Moosavi. 2025. [Beyond hate speech: NLP’s challenges and opportunities in uncovering dehumanizing language](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26965–26980, Suzhou, China. Association for Computational Linguistics.
- Severin Schweitzer and Adam Waytz. 2021. [Language as a window into mind perception: How mental state language differentiates body and mind, human and nonhuman, and the self from others](#). *Journal of Experimental Psychology: General*, 150(8):1642–1672. Epub 2020 Dec 14.
- Jeroen Vaes, Paola Paladino, and Elisa Puvia. 2011. [Are sexualized women complete human beings? why men and women dehumanize sexually objectified women](#). *European Journal of Social Psychology*, 41(6):774–785.
- Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. [Detecting East Asian prejudice on social media](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.
- Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. [Cross-replication reliability - an empirical approach to interpreting inter-rater reliability](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [Predicting the type and target of offensive posts in social media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

## A Operationalizing the Dual Model of Dehumanization

Table 2: Examples illustrating the distinction between offensive language, hate speech, and dehumanization (hereafter dehum). The examples are informed by our data and highlight that these constructs are conceptually and empirically distinct, although overlaps occur in practice.

| Category                                  | Example                                      | Interpretation                                                      |
|-------------------------------------------|----------------------------------------------|---------------------------------------------------------------------|
| Offensive (non-dehum)                     | All assholes here. Damn!                     | Offensive, but untargeted and no denial of humanness                |
| Hate speech (non-dehum)                   | He is right! We should despise group X.      | Targets a group, is hateful but not dehumanizing                    |
| Hate speech (non-dehum)                   | Group X should not be allowed to live here.  | Targets a group, is discriminatory but not dehumanizing             |
| Expl. animalistic dehum (hate-speech)     | These people look and behave like rats.      | Direct comparison to animals, indirect group reference, hate-speech |
| Expl. Mechanistic dehum (non-hate-speech) | She is a bot not a human.                    | Direct comparison with non-living object, no group reference        |
| Impl. animalistic dehum (hate-speech)     | Group X is incapable of thinking rationally. | Denial of human trait, group reference, discriminatory.             |
| Counter-speech (non-dehum)                | Stop calling them animals, that is wrong.    | Contains dehumanizing terms but rejects them                        |

## B Target Selection

| Target                  | Terms                                                                                                                                                                                                                                                                                                                                          |
|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Foreigners / Immigrants | immigrant, immigrants, refugee, refugees, alien, aliens, illegal, illegals, asylum, escapee, escapees, foreigner, foreigners, exile, displaced person, displaced persons, migrant, migrants, undesirables                                                                                                                                      |
| Religion                | jew, jews, hebrew, hebrews, kike, kikes, yid, yids, mockie, mockies, mocky, mockys                                                                                                                                                                                                                                                             |
| Race                    | black, blacks, bootlip, bootlips, dindu, dindus, porch monkey, porch monkeys, ghetto monkey, ghetto monkeys, golliwog, golliwogs, hoodrat, hoodrats, jigga, jigger, jigro, kneegrow, kneegrows, moulie, moulies, negro, negros, nigger, niggers, afro-american, afro-americanjns, afromerican, afromericans, native-american, native-americans |
| Gender / Sex / Identity | woman, women, girl, girls, bitch, bitches, cunt, cunts, dyke, dykes, pussy, pussies, hoe, hoes, bint, bints, female, females, shemale, shemales, trans, twat, twats, asexual, asexuals, faggot, faggots, homosexual, homosexuals                                                                                                               |
| Ethnicity / Nationality | japanese, german, germans, american, americans, asian, asians, turk, turks, arab, arabs, chinese, mexican, mexicans, african, africans, indian, indians, italian, italians, spanish, russian, russians, scot, scots, french, syrian, syrians, dutch                                                                                            |
| Political               | democrat, republican, republicans, democrats, dems, reps                                                                                                                                                                                                                                                                                       |
| Disability              | cripple, cripples, gimp, gimps, libtard, libtards, mongs, mongos, retard, spaz, tard, tarded, tards, mongo, mongos                                                                                                                                                                                                                             |
| Old people              | elderly, elderlies, olderwoman, olderwomans, olderman, oldermen, oldster, oldsters, oldppl, old person, old persons                                                                                                                                                                                                                            |
| Medical and criminal    | patient, patients, criminal, criminals, inmate, inmates, prisoner, prisoners                                                                                                                                                                                                                                                                   |

Table 3: List of frequently dehumanized target groups in English.

| Target                    | Terms                                                                                                                                                                                                                                                                                                      |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Foreigners / Immigrants   | Einwanderer, Immigrant, Immigranten, Flüchtling, Flüchtlinge, Ausländer, Illegale, Illegaler, Vertriebener, Vertriebene, Unerwünschte, Migrant, Migranten                                                                                                                                                  |
| Religion                  | Jude, Juden, Hebräer, Jews, Jew, Israeli, Israelis, Palästinenser                                                                                                                                                                                                                                          |
| Race/Ethnicity            | Schwarzer, Schwarze, Neger, Nigger, Niggers, Afroamerikaner, Ureinwohner Amerikas, Chines, Asiate, Asiaten, Araber, Türke, Türken, Mexikaner, Afrikaner, Inder, Italiener, Spanier, Russe, Russen, Schotte, Schotten, Franzose, Franzosen, Syrer, Niederländer, Ukrainer                                   |
| Gender / Sex / Identity   | Frau, Frauen, Mädchen, Schlampe, Schlampen, Fotze, Fotzen, Lesbe, Lesben, Muschi, Muschis, Hure, Huren, Tussi, Tussis, Weiblich, Weibliche, Transe, Transen, Trans, Transsexuell, Transsexuelle, LGBTQ, Asexuell, Asexuelle, Schwuchtel, Schwuchteln, Homosexuell, Homosexuelle, Schwule, Schwule Menschen |
| Political                 | Republikaner, Konservative, Demokrat, Demokraten, Grüne, Ökos, Grünlinge, Sozialdemokraten, SPDler, SPD, CDUler, Christdemokraten, CSUler, CSU, AfDler, Blaue, Rechtskonservative, Liberale, FDPler, Gelbe, Linke, Politiker                                                                               |
| Disability                | Krüppel, Behinderter, Behinderte, Libtard, Libtards, Mongo, Mongos, Spasti, Spasten, Tard, Tards, Behindert                                                                                                                                                                                                |
| Old people                | Ältere/r, Ältere, Ältere Frau, Ältere Frauen, Älterer Mann, Ältere Männer, Senior, Senioren, Alte Leute, Alte Person, Alte Personen                                                                                                                                                                        |
| Medical and criminal      | Patient, Patienten, Krimineller, Kriminelle, Diebe, Straftäter, Häftling, Häftlinge, Gefangener, Gefangene                                                                                                                                                                                                 |
| Social class / Occupation | Arbeiter, 9-bis-5-Arbeiter, 9-to-5, Tourist, Reisende, Geschäftsmann, Geschäftsmänner                                                                                                                                                                                                                      |

Table 4: List of frequently dehumanized target groups in German.

## C Annotation Task and Guidelines

In the following, we describe the annotation task and the guidelines provided to crowd workers for the different dehumanization annotation tasks. Prior to starting the task, participants were informed that they would be exposed to potentially harmful or offensive content and that they could discontinue the annotation at any time. Also they were informed how their annotations would be used only by researchers. Annotators were compensated in accordance with the German minimum wage.

### C.1 Concrete Task Example

All Democrats are toxic. Only difference is the rates that they parasitically infect new hosts.

**Would you consider this post as being offensive towards a person or a group?**

☐ yes

☐ no

☐ can't tell

**Please highlight any parts of the text where a person or group is negatively compared to animals, inferior creatures, diseases or to inanimate objects. Only highlight examples that reinforce or endorse these negative comparisons. Otherwise choose 'no such comparison'.**

☐ Animal Comparison (e.g. cockroaches, rats)

☐ Inf. Creature Comparison (e.g. monsters, barbarians)

☒ Disease Comparison (e.g. virus, pest)

☐ Inanimate Comparison (e.g. objects, machines)

☐ No such comparison

**Please highlight the target of any negative comparison in the text, but only if such a comparison exists.**

☒ Target

☐ No Target

**How confident are you that your evaluation of the text is correct?**

☐ Very sure

☐ Sure

☐ Unsure

☐ Very unsure

Keyboard Input:

|               |              |                                            |                                                      |                                       |                                               |        |
|---------------|--------------|--------------------------------------------|------------------------------------------------------|---------------------------------------|-----------------------------------------------|--------|
| ←             | →            | 1                                          | 2                                                    | 3                                     | 4                                             | 5      |
| Move backward | Move forward | Animal Comparison (e.g. cockroaches, rats) | Inf. Creature Comparison (e.g. monsters, barbarians) | Disease Comparison (e.g. virus, pest) | Inanimate Comparison (e.g. objects, machines) | Target |

Move backward
Move forward

Figure 4: An example of the annotation task that was presented to the crowd-worker (here explicit dehumanization).

## C.2 Explicit Dehumanization Instructions

**Explicit Dehumanization Task** Please read the entire task description and examples carefully! Your task involves multiple steps.

- First, we would like to know whether you perceive a text as generally **offensive** toward a person or a group of people.
- Second, we ask you to find and highlight when the text **negatively compares** a person or group to **animals** (e.g., cockroaches, dogs), **inferior creatures** (e.g., monsters, savages), **diseases** (e.g., viruses, pests), or **inanimate objects** (e.g., machines, tools, or robots). We do not distinguish between direct or relative comparisons, so if one group is compared in relation to another group, mark it. Focus only on the text, and ignore any hyperlinks.

Only highlight examples that **reinforce** or **endorse** these **negative comparisons**. Do not highlight statements that argue against them. Besides, not all comparisons are harmful—for example, saying someone is “strong like a lion” is not dangerous, so do not mark those.

Please highlight both the **negative comparison** (if present) and the **target**, including any pronouns referring to the target. Examples are provided below.

### Direct Comparison

- That’s the thing with the **Jews**. They’re just **like monkeys** with their lack of knowledge.  
*Explanation: This statement directly associates Jewish people with monkeys.*
- **Women** = kitchen, kitchen = dishes, dishes = objects, **objects** = **women**.  
*Explanation: This statement reduces women to being associated solely with the kitchen and household chores, ultimately equating them with objects.*
- **Republicans** are hateful **ugly creatures**!  
*Explanation: The author of the post directly compares members of the Republican party to creatures.*
- Those **curry boys** are **filthy little pigs**, so gross.  
*Explanation: The statement directly compares a group of people to pigs.*

### Indirect Comparison through Metaphorical Language and Analogies

- **A wild flock** of **Jews** quickly **flies away**.  
*Explanation: Jewish people are implicitly compared to birds through the metaphor of a flock, typically used to describe a group of birds.*
- **She** is my **possession**. I can use **her** as I like.  
*Explanation: This statement implies that women are considered belongings or property. The metaphorical language used here indirectly associates women with objects that can be owned and controlled, thereby dehumanizing them and reducing their value to that of an item rather than a person.*

### No (Endorsed) Negative Comparison

- He is strong like a lion.  
*Explanation: The statement does not demean the person being compared. Instead, it uses a positive comparison to highlight strength, a trait commonly associated with lions. While it likens someone to an animal, it does so in a way that is complimentary rather than degrading, and therefore does not lower social barriers.*
- Who the fuck do you think you are calling migrants cockroaches? Btw, I’m one of those migrants.  
*Explanation: The author of the post is not expressing the opinion that migrants are cockroaches. Instead, they are challenging and rejecting this comparison, making it clear that they do not agree with or accept this insult.*

## No Comparison

- Women. Aren't. Animals.

*Explanation: This statement serves as counterspeech, explicitly rejecting the idea that women should be compared to animals.*

- Women love cats.

*Explanation: This is a general statement expressing a common stereotype that some women like cats.*

## C.3 Implicit Animalistic Dehumanization Instructions

**Implicit Animalistic Task** Please read the entire task description and examples carefully! Your task involves multiple steps.

- First, we would like to know whether you perceive a text as generally **offensive** toward a person or a group of people.
- Second, we ask you to find and highlight any part of the text that directly or indirectly **portrays** an **individual** or a **group of humans** as being **incivil/uncultured** (rude, coarse, disrespectful, or lacking in courtesy and politeness), **irrational** (actions, thoughts, or beliefs that contradict reason or logic), **unintelligent** (lack of mental ability, understanding and difficulty grasping complex ideas), **immature/child-like** (lack of emotional development, responsibility, or maturity), or **lacking morals** (dishonest, unethical, or unscrupulous, prone to violate the moral code). Focus only on the text, and ignore any hyperlinks.

Please highlight both the **comparison** (if present) and the **target**, including any pronouns referring to the target. Examples are provided below.

### Highlights

- Some **women** these days **have no morals** whatsoever.  
*Explanation: Women are explicitly depicted as **lacking morals**.*
- Most **Asians** have **no manners**; it's like they are **allergic to saying thank you**.  
*Explanation: The author claims that Asian people are **uncultured**.*
- **Women** are actually capable of logical thinking, perhaps **not to the same degree as men**.  
*Explanation: Women are described as being **irrational**, compared to men.*
- **They** are thinking outside the box, but unfortunately, **they're not thinking at all**.  
*Explanation: The language used suggests that individuals are being characterized as **unintelligent**.*
- Watching **Republicans** in Congress feels like **babysitting a room full of toddlers**. **Tantrums, refusing to share, and making a mess**.  
*Explanation: Republicans are portrayed as **immature** and behaving like toddlers.*

### No Highlights

- When people say Arabs are rude, I just get a little tired of them.  
*Explanation: It is not the opinion of the author that people from a certain ethnic group are rude.*
- Irrationality is a curse.  
*Explanation: No target group or individual is accused of being irrational.*
- I hate these fucking women bitches!  
*Explanation: While this sentence is derogatory towards women, the target group is not described as incivil/uncultured, coarse, irrational, unintelligent, immature/child-like, or lacking morals.*

## C.4 Implicit Mechanistic Dehumanization

**Implicit Mechanistic Instructions** Please read the entire task description and examples carefully! Your task involves multiple steps.

- First, we would like to know whether you perceive a text as generally **offensive** toward a person or a group of people.
- Second, we ask you to find and highlight any part of the text that directly or indirectly **portrays** an **individual** or a **group of humans** as being **cold** (emotionally distant, unfeeling, or lacking warmth and compassion), **passive/lacking agency** (accepting or allowing things to happen without active response or resistance, often showing a lack of initiative or assertiveness), **superficial/lacking depth** (concerned only with surface appearances or shallow aspects, lacking depth, substance, or thoroughness), **fungible/interchangeable** (someone who is interchangeable or replaceable, typically without any loss of value), or **rigid/inflexible** (someone who is inflexible, unwilling to change, or firmly fixed in position, behavior, or beliefs, often lacking adaptability). Focus only on the text, and ignore any hyperlinks.

Please highlight both the **comparison** (if present) and the **target**, including any pronouns referring to the target. Examples are provided below.

### Highlights

- Trying to get compassion from **Republicans** is like **asking a thermostat for a hug**.  
*Explanation: Republicans are described as **emotionally distant**.*
- Some **people** have **zero backbone**. Just **nodding along**, **agreeing with whatever**, **like a human echo**.  
*Explanation: People are portrayed as having no **agency**.*
- Some **men** got a personality as **deep as a puddle**—looks nice for a second, then you realize **there's nothing there**.  
*Explanation: Men are described as being **superficial** beings.*
- There's **nothing unique** about **Mexicans**. Swap them out with a dozen others, and **no one would notice**.  
*Explanation: The language used suggests that Mexicans are being characterized as **interchangeable**.*
- Talking to **old people** is **like trying to reprogram a 20-year-old computer—impossible** and not worth the effort.  
*Explanation: Old people are portrayed as **rigid** and set in their own ways without willingness to change.*

### No Highlights

- When people say Arabs are shallow, I just get a little tired of them.  
*Explanation: It is not the opinion of the author that people from a certain ethnic group are superficial.*
- Passivity is the enemy of progress.  
*Explanation: No target group or individual is accused of being passive.*
- I hate these fucking women bitches!  
*Explanation: While this sentence is derogatory and thus **offensive** towards women, the target group is not described as passive, superficial, fungible, or rigid. Thus, it should be marked as offensive but there should be no highlights.*

## C.5 Annotation Procedure Diagram

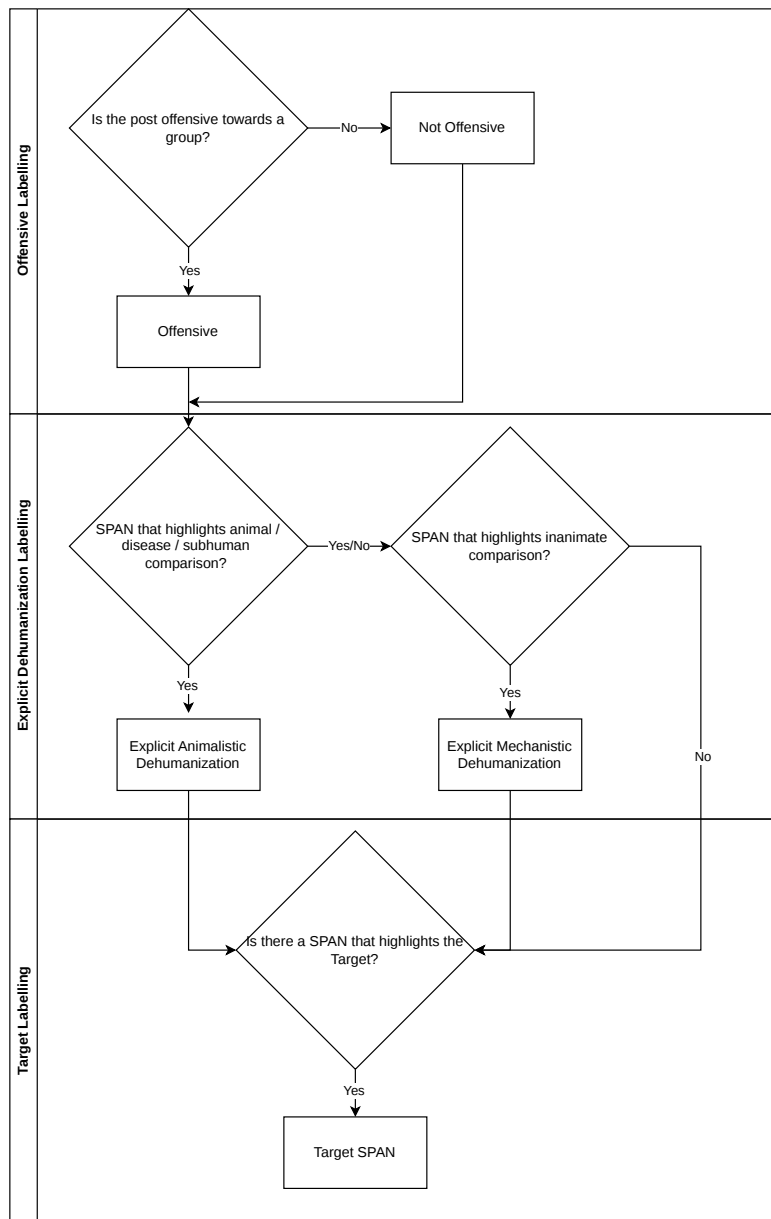


Figure 5: Annotation Schema for the Explicit Dehumanization Task. The annotation procedure for the implicit dehumanization task is analogous, with the only difference being the span labeling of the respective implicit dehumanization subclasses.

## D Most Frequent Targets

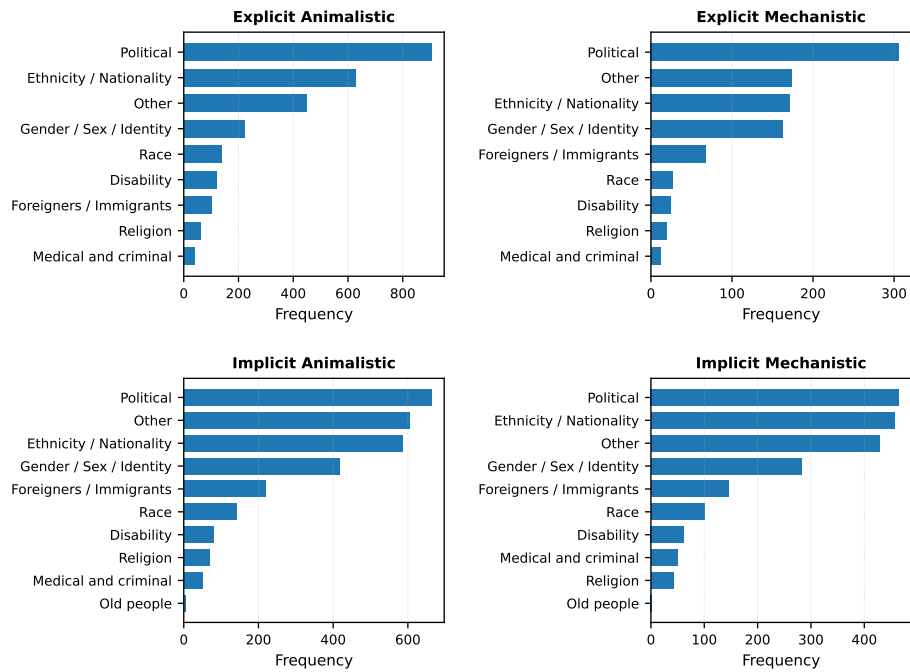


Figure 6: Overview of target groups in English instances labeled as dehumanizing by crowdworkers/experts

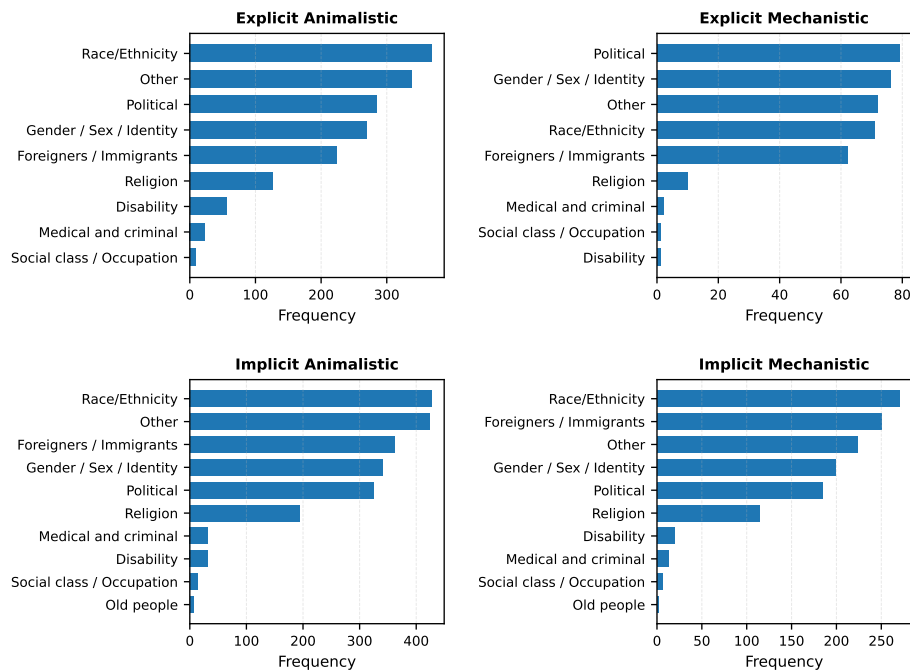


Figure 7: Overview of target groups in German instances labeled as dehumanizing by crowdworkers/experts

## E Candidate Selection Methods

### E.1 Theory-informed Instance Selection:

We utilize a theory-informed selection method based on survey-scale items designed to measure various dimensions of dehumanization (Rasset et al., 2023; Kuljian and Hohman, 2023; Haslam et al., 2008). These survey items, combined with the previously identified frequent targets of dehumanization, are used to create template instances. These templates serve as a reference to find semantically similar instances within our real-world datasets from the two platforms. For example, a survey item proposed by Kuljian and Hohman for measuring implicit animalistic dehumanization is:

[TARGETS] are childish in their decision making

Here, the target is highlighted in blue, and the denial of the human trait—in this case, maturity—is highlighted in red, capturing a subtle form of animalistic dehumanization. Using this approach, we systematically vary the targets and the denial of human traits across various dehumanization dimensions. We use three different SBERT models (Reimers and Gurevych, 2019) to identify real Twitter or Reddit posts that are semantically similar to our predefined templates. For each instance in our dataset, we generate a corresponding sentence embedding and compute its cosine similarity with all template embeddings, ranking instances by their similarity scores for each model.

### E.2 Dataset-Informed Instance Selection

As a second strategy for instance selection, we utilize existing dehumanization datasets to identify semantically similar examples within our corpus. Specifically, we collect all dehumanizing instances from publicly available English datasets (Kirk et al., 2023; Vidgen et al., 2021, 2020; Sachdeva et al., 2022) and identify the three nearest neighbors (again using SBERT based embedding similarities) in our corpus, explicitly excluding exact duplicates. Since existing datasets focus mainly on explicit animalistic dehumanization dimension we expect to find instances that also reflect this dimension.

### E.3 LLM-Informed Instance Selection

In our third instance selection approach, we utilize Llama 3.1-70B-Instruct to generate artificial instances by prompting the model to create realistic examples for each dimension of dehumanization (2756 English and 3300 German). We experiment with various strategies to bypass the model’s guardrails that prevent it from producing harmful content. We use these synthetic examples solely for information retrieval, aiming to again identify nearest-neighbor real-world posts within our dataset using SBERT.

### E.4 Candidate Selection Results

Table 5 reveals for the different candidate selection strategies the fraction of candidates that were labeled as dehumanization by the majority of annotators. Note that we have not used existing datasets to select implicit candidates since they only focus on the explicit dehumanization dimension. Sampling methods differ in their ability to find dehumanizing examples in the four relevant categories. Theory-informed candidate selection provides the highest fraction of positive instances for all categories, except for implicit mechanistic dehumanization. Here, LLM-based templates are more efficient in finding relevant instances in the real-world data.

### E.5 Diversity Analysis

We examined the diversity of instances positively annotated as dehumanization, originating from different candidate selection approaches. We calculated the pairwise cosine distances between their sentence embeddings. For the explicit dimension, each candidate selection strategy produced a relatively diverse set of candidates, with an average intra-strategy cosine distance of 0.79 and 0.80 for implicit instances. A visual representation of the sentence embeddings, projected into a two-dimensional space using UMAP, supports the finding that all sampling strategies find relatively diverse set of instances (cf. Figure 8).

| Category   | Theory | LLM | Existing |
|------------|--------|-----|----------|
| Exp Animal | 54%    | 21% | 16%      |
| Exp Mecha  | 22%    | 8%  | 5%       |
| Imp Animal | 75%    | 54% | -        |
| Imp Mechan | 40%    | 65% | -        |

Table 5: Fraction of dehumanizing instances retrieved with a sampling method (column) for a specific dimension of dehumanization (row). This table shows that theory-informed instance selection leads to the highest fraction of dehumanizing instances across almost all four dimensions of dehumanization. Only for the implicit mechanistic dimensions, LLM-based instance selection leads to even more positive instances.

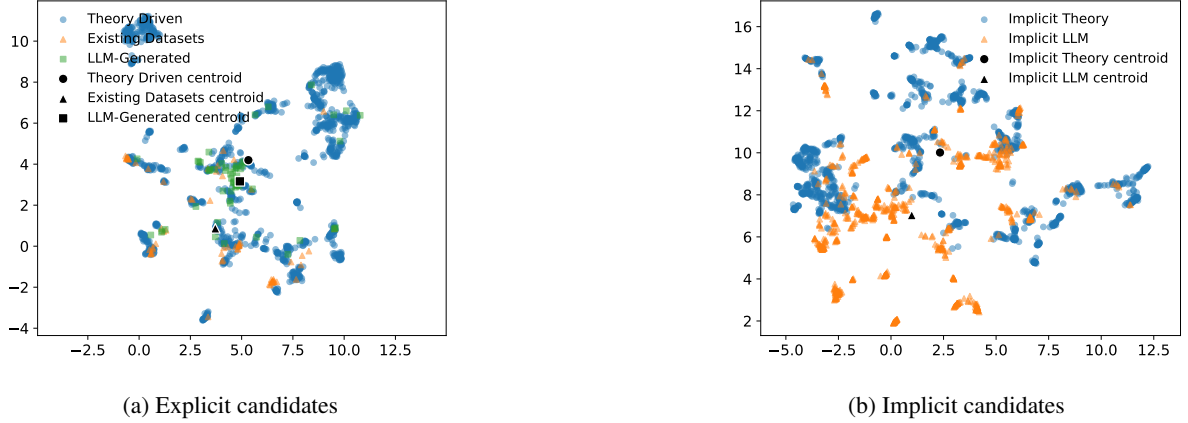


Figure 8: Semantic space of (a) explicit and (b) implicit candidates that are positively labeled by the crowd as dehumanization.

## F Experiments

### F.1 German Experiments Results

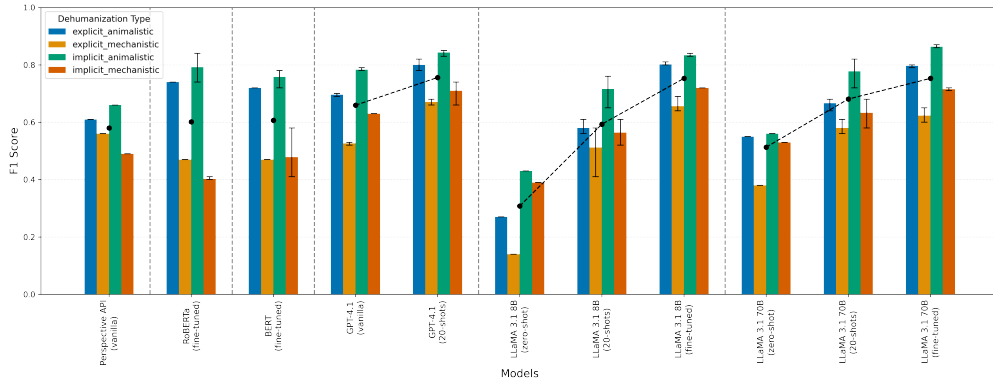


Figure 9: Test  $F_1$ -Scores for the four different dimensions of dehumanization (see color bars) on German Reddit and Twitter data for different model architectures. For the German data we used multilingual versions of BERT and RoBERTa. HateBERT and MetaHateBERT are not tested as they are pretrained and fine-tuned on English data.

### F.2 Animalistic vs. Mechanistic

### F.3 Implicit vs. Explicit

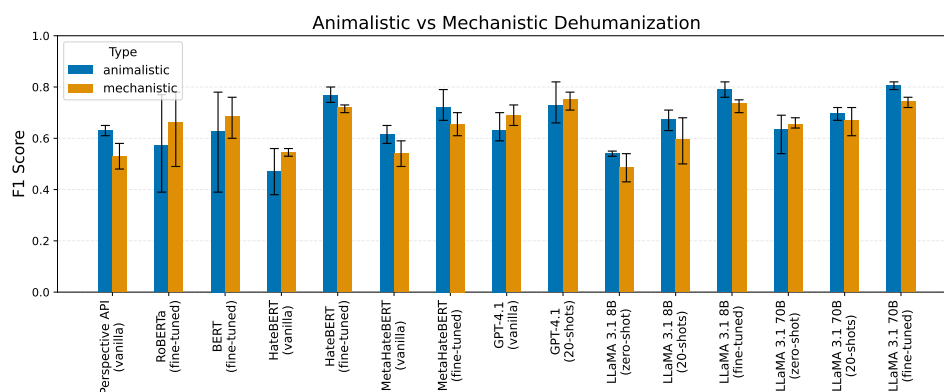


Figure 10: Model comparison: Animalistic vs. Mechanistic Dehumanization (English)

## G Candidate Data

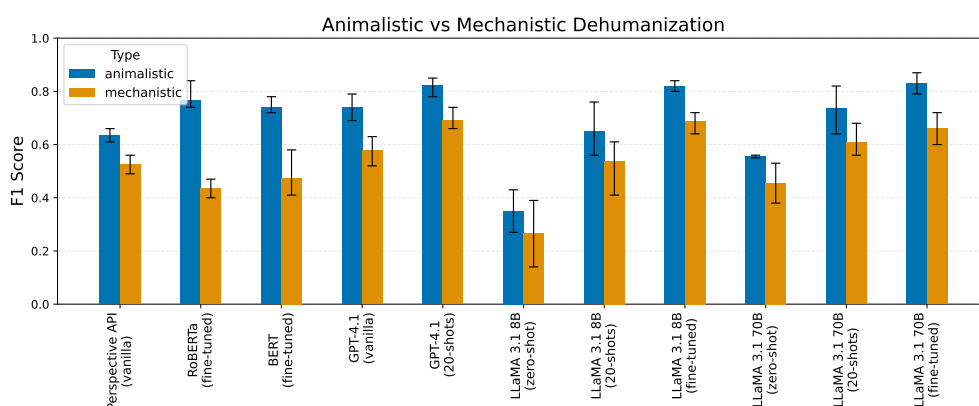


Figure 11: Model comparison: Animalistic vs. Mechanistic Dehumanization (German). HateBERT and MetaHateBERT

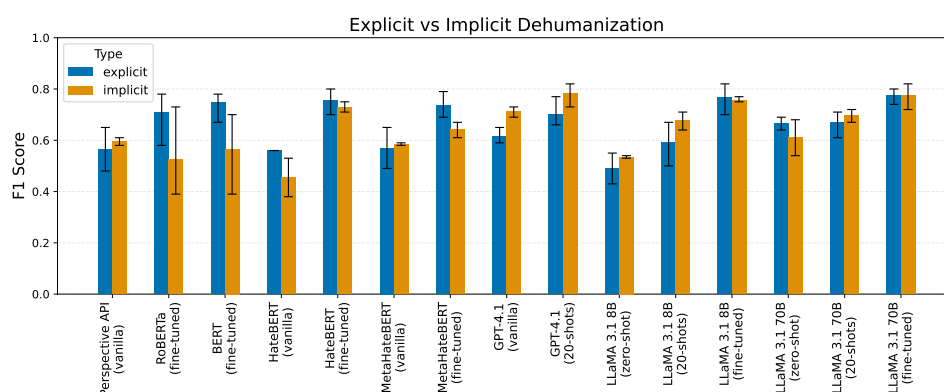


Figure 12: Model comparison: Explicit vs. Implicit Dehumanization (English)

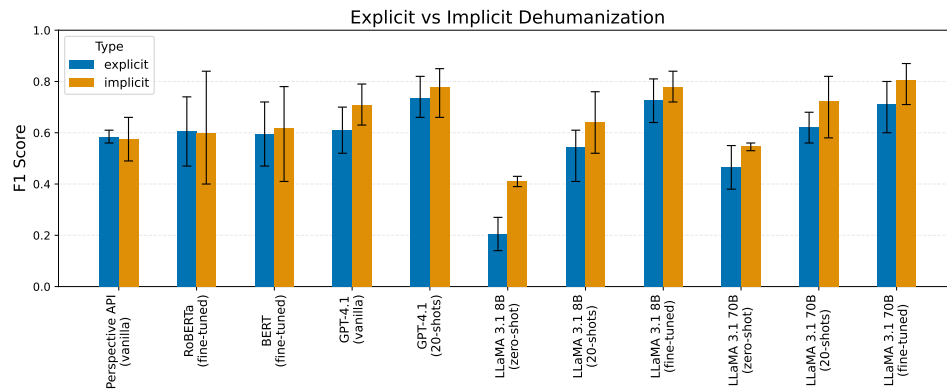


Figure 13: Model comparison: Explicit vs. Implicit Dehumanization (German)

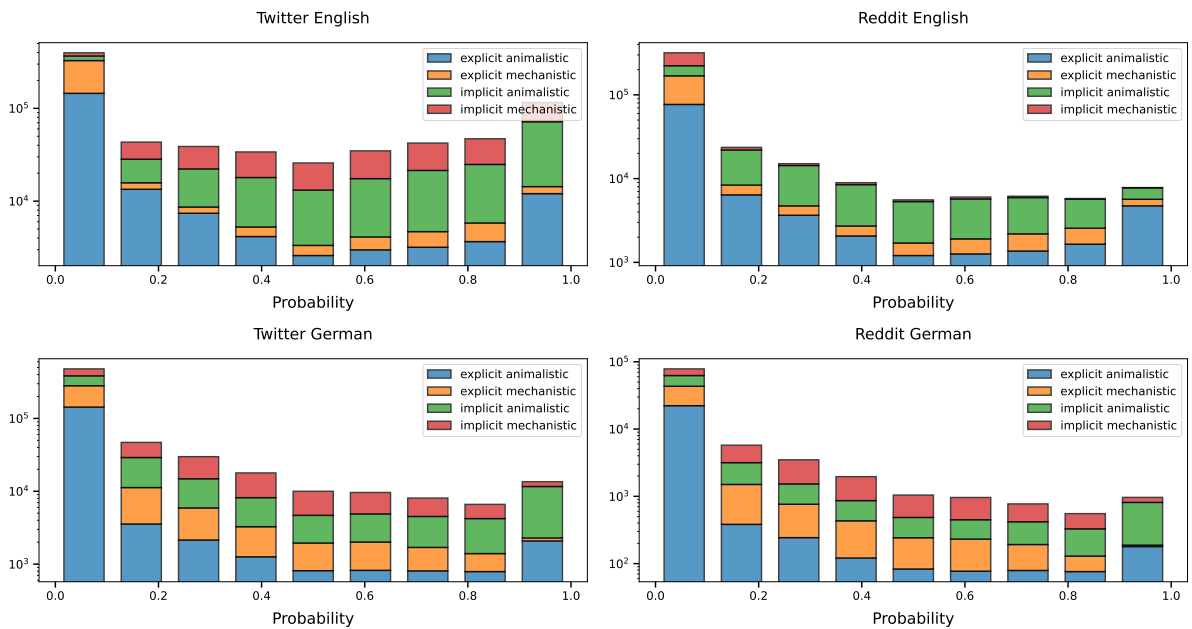


Figure 14: Distribution of predicted logit probabilities for the positive class as produced by the best-performing fine-tuned Llama 3.1 8B models across our candidate datasets in both languages.

## H Annotator demographics

|    | Dataset     | #Ann. | Age M (SD)    | Time M (SD)   |
|----|-------------|-------|---------------|---------------|
| EN | Explicit    | 400   | 38.36 (12.54) | 25.26 (13.47) |
|    | Implicit    |       |               |               |
|    | Animalistic | 200   | 37.22 (12.71) | 27.71 (14.98) |
|    | Mechanistic | 200   | 37.03 (11.54) | 26.53 (12.80) |
|    | Total       | 741   | 37.10 (12.47) | 30.44 (15.89) |
| DE | Explicit    | 223   | 35.09 (11.44) | 22.32 (10.30) |
|    | Implicit    |       |               |               |
|    | Animalistic | 151   | 34.84 (11.46) | 19.34 (10.56) |
|    | Mechanistic | 147   | 33.97 (11.05) | 18.56 (9.43)  |
|    | Total       | 292   | 33.92 (10.78) | 22.32 (10.67) |

Table 6: Prolific annotator statistics. EN represents the English dataset and DE the German. #Ann. refers to the number of annotators. Age M (SD) is the mean age and standard deviation of annotators. Time M (SD) denotes the mean time spent (in minutes) with standard deviation. Note that the total number of annotators is less than the sum of all groups because some annotators worked on multiple datasets.

|    | Dataset  | Men (%)     | Women (%)   | Other (%) |
|----|----------|-------------|-------------|-----------|
| EN | Explicit | 194 (47.74) | 200 (50.25) | 4 (1)     |
|    | Implicit |             |             |           |
|    | Anim.    | 106 (53.00) | 88 (44.00)  | 6 (3.00)  |
|    | Mech.    | 102 (51.00) | 94 (47.00)  | 4 (2.00)  |
|    | Total    | 364 (65.82) | 366 (411)   | 10 (1.25) |
| DE | Explicit | 275 (64.49) | 141 (33.25) | 8 (1.8)   |
|    | Implicit |             |             |           |
|    | Anim.    | 87 (57.61)  | 58 (38.41)  | 6 (3.97)  |
|    | Mech.    | 89 (60.54)  | 54 (36.73)  | 4 (2.72)  |
|    | Total    | 171 (58.76) | 112 (38.48) | 8 (2.74)  |

Table 7: Distribution of gender among Prolific annotators. Because of the low number of German annotators we allowed individual crowd-workers to annotate one dimension multiple-times (max 3).

| Dataset  | HS (%)      | UG (%)      | Grad (%)    | PhD (%)   | N/A (%)  |
|----------|-------------|-------------|-------------|-----------|----------|
| Explicit | 42 (21.00)  | 100 (50.00) | 51 (25.50)  | 6 (3.0)   | 1 (0.50) |
| Implicit |             |             |             |           |          |
| Anim.    | 51 (25.50)  | 95 (47.50)  | 43 (21.50)  | 9 (4.50)  | 2 (1.00) |
| Mech.    | 47 (23.50)  | 106 (53.00) | 42 (21.00)  | 4 (2.00)  | 1 (0.50) |
| Total    | 122 (22.06) | 279 (50.45) | 130 (23.51) | 18 (3.25) | 4 (0.72) |

Table 8: Distribution of education level among Prolific annotators, for the English dataset. *HS* refers to high school, *UG* to undergraduate, *Grad* to graduate, and *PhD* to doctoral education. *N/A* indicates missing or unreported data. German annotators did not report their education level.

| Dataset  | W (%)       | B (%)      | A (%)     | O (%)     | N/A (%)  |
|----------|-------------|------------|-----------|-----------|----------|
| Explicit | 300 (75.37) | 64 (16.08) | 20 (5.02) | 14 (3.51) | 1 (0.50) |
| Implicit |             |            |           |           |          |
| Anim.    | 149 (74.5)  | 22 (11.0)  | 23 (11.5) | 4 (2.0)   | 2 (1.00) |
| Mech.    | 152 (76.0)  | 19 (9.5)   | 18 (9.0)  | 10 (5.00) | 1 (0.50) |
| Total    | 559 (75.54) | 98 (13.24) | 56 (7.56) | 27 (3.64) | 4 (0.72) |

Table 9: Distribution of race among Prolific annotators, for the English dataset. *W* refers to White, *B* to Black, *A* to Asian, and *O* to Other. *N/A* indicates missing or unreported data. This question was not included for the German dataset. Individual annotators can be present in different dimensions.

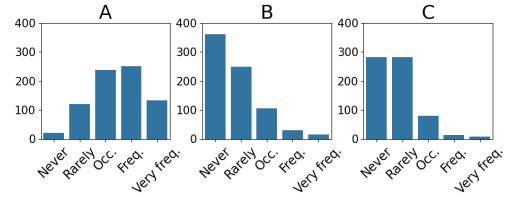


Figure 15: Frequency with which the annotators of the English part of the dataset (A) encounter dehumanizing content online (B) are the target of dehumanizing content on social media and (C) are the target of dehumanizing content outside of social media

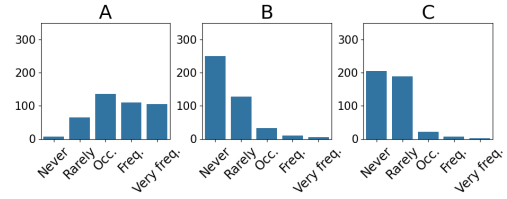


Figure 16: Frequency with which the annotators of the German part of the dataset (A) encounter dehumanizing content online (B) are the target of dehumanizing content on social media and (C) are the target of dehumanizing content outside of social media

## I Inter Annotator Agreement

| Dataset       | English | German |
|---------------|---------|--------|
| Explicit      | 0.393   | 0.475  |
| Animalistic   | 0.431   | 0.518  |
| Animal        | 0.577   | 0.702  |
| Subhuman      | 0.275   | 0.358  |
| Disease       | 0.459   | 0.468  |
| Mechanistic   | 0.381   | 0.394  |
| Offensive     | 0.425   | 0.328  |
| Implicit      | 0.286   | 0.436  |
| Animalistic   | 0.286   | 0.517  |
| Irrational    | 0.124   | 0.165  |
| Morals        | 0.333   | 0.366  |
| Unintelligent | 0.392   | 0.693  |
| Child-like    | 0.267   | 0.284  |
| Incivil       | 0.230   | 0.220  |
| Mechanistic   | 0.290   | 0.285  |
| Superficial   | 0.243   | 0.125  |
| Cold          | 0.155   | 0.255  |
| Passive       | 0.151   | 0.269  |
| Fungible      | 0.102   | 0.142  |
| Rigid         | 0.142   | 0.295  |
| Offensive     | 0.321   | 0.330  |

Table 10: IAA agreement for the different categories.

## J Crosslingual Analysis

In addition to our monolingual analysis, we evaluated the performance of Llama 3.1 different in cross-lingual settings. Specifically, we tested whether models fine-tuned on data from one language can generalize to another. Figure 17 presents results for the explicit animalistic and mechanistic dehumanization dimensions. The findings demonstrate that our annotated data enables robust model performance even in cross-lingual applications, suggesting generalizability across linguistic and cultural boundaries. Notably, models fine-tuned on a different language (represented by hatched bars) performed comparably to models fine-tuned on the evaluation language and outperformed few-shot baselines. Interestingly, the explicit mechanistic model fine-tuned on English data outperforms the model fine-tuned on German data when evaluated on the German test set. We attribute this to the smaller number of positive training instances available in the German dataset compared to the English dataset (see Table 1).

## K AI Usage Acknowledgement

We used ChatGPT (OpenAI) for language editing purposes, including improvements to grammar, wording, and overall readability of the manuscript.

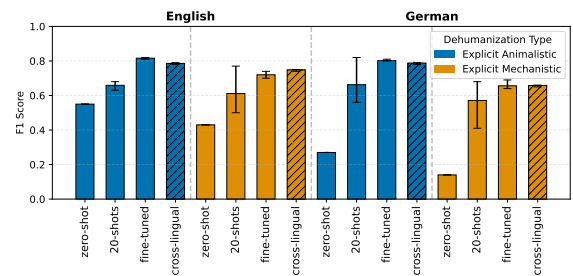


Figure 17: Cross-lingual analysis of explicit dehumanization models. The labels on top indicate the evaluation language. Cross-lingual models (hatched bars) were fine-tuned on the opposite language.