Motivation for Dataset Creation

Why was the dataset created?

The dataset was created to enable research on non-negative abusive stereotypes and its semantic variants about identity groups, particular its detection. Previous datasets for abusive language detection do not sufficiently cover this form of abuse.

What (other) tasks could the dataset be used for?

While in our paper we primarily focused on the *detection* of non-negative stereotypes on identity groups, we also envisage its use for evaluating the impact of varying degrees of generalization and perspective framing or the dependence of the abuse on the identity group.

Has the dataset been used for any tasks already?

No.

Who funded the creation of the dataset?

It was funded by the first author's institution.

Dataset Composition

What are the instances?

The instances are constructed. Each instance represents a sentence. The instances are abusive, not abusive or ties, representing either potential stereotypes or prevalent characteristics targeting specific identity groups (i.e. *Black people, gay people, Jews, Muslims, women*). Each of these subsets consists of atomic sentences and eight semantic variants (i.e. *implication, "always", "all", "many", "some", instantiation, self-identification, authoritative report*). For those atomic instances that result in a meaningful sentence, the identity group was switched and substituted by any of the other 4 groups.

Are relationships between instances made explicit in the data?

No. The dataset was sampled from a wide set of different crowdworkers (i.e. more than 350, approximately 80 crowdworkers per identity group) in order to avoid an author bias. In general, our dataset does not provide any information on the crowdworkers. Therefore, this dataset is hardly suitable for studying the relationships between instances.

How many instances of each type are there?

The dataset comprises 11,204 English sentences in total, where 4,122 sentences are abusive, 6,967 sentences are non-abusive sentences and the remaining 115 sentences are ties.

What data does each instance consist of?

For each instance, we provide the following information:

- The main dataset:
    - The atomic sentence itself.
    - The identity group (i.e. *Black people, gay people, Jews, Muslims, women*).
    - The corresponding semantic variants (i.e. *implication, "always", "all", "many", "some", instantiation, self-identification, authoritative report*).
    - The binary class label indicating whether the instance was rated as abusive or not. This label has been established via crowdsourcing, i.e. it is the result of manual annotation. Each label represents the majority label over ratings provided by 5 different crowdworkers. A very small number of the labels are ties.
    - For each sentence we also include the manually extracted features. These are features that emerged from the data and were created by the co-authors of the paper. Unlike the subset, the topics are not mutually exclusive:
        - subset
            - potential stereotype
            - prevalent characteristic
        - topic
            - outward_appearance
            - character_traits/habit
            - preferences
            - work/education
            - food/drink
            - cultural/historical_characteristics
            - social_interaction
            - living_conditions
            - other
            - physiology/biology/medicine
            - competence
            - religious_practice
            - sex
            - family
- The extended component of the dataset:
    - The meaningful, replaced atomic sentences of the switched identity groups
    - The binary class label of the atomic instances and any other of the switched identity groups indicating whether the instance was rated as abusive or not. This label has also been established via crowdsourcing, i.e. it is the result of manual annotation. Each label represents the majority label over ratings provided by 5 different crowdworkers. A very small number of the labels are ties.
    - For each sentence we have added the feature 'group' to mark to which identity group each atomic original sentence belongs.

Is everything included or does the data rely on external resources?

Everything is included.

Are there recommended data splits or evaluation measures?

As the most unbiased set-up, we recommended to arrange the data splits in test and training data. The release of the dataset includes such partitioning for the 5-fold cross-validation as it was used for the supervised classification experiments in the paper.

What experiments were initially run on this dataset?

We examined different forms of classification, stereotype classification, cross-group classification, within-group classification, atomic classifier, GPT-4 and human classifier.

For supervised learning we considered the transformer DeBERTa and fine-tuned the pretrained model on the given training data using the FLAIR-framework. We always report the average over five training runs (+ standard deviation).

We examined two publicly available classifiers for abusive language detection. We use PerspectiveAPI, i.e. HateBERT fine-tuned on ToxiGen. We also fine-tuned a transformer on the ISHate.

We fine-tuned a classifier on each of two recent datasets for stereotype classification. As datasets, we chose StereoSet and the dataset by Pujari et al. (2022) since they contain complete sentences including negative data.

As cross-group classification, we considered a classifier trained on four identity groups of our dataset, testing it on the remaining one.

As within-group classification, we considered a classifier trained on the instances of the same identity group using a 5-fold crossvalidation.

We considered two versions (i.e., oracle and auto) of the atomic classifier and also two types (i.e., zero-shot and an augmented version) of GPT-4. For the latter, we are augmenting each sentence in our dataset with the respective completions obtained from the zero-shot approach.

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

Data Collection Process

How was the data collected?

The dataset was not collected from existing sources. Instead, crowdworkers of five identity groups who were required to be native speakers of English (i.e., *Black people, gay people, Jews, Muslims, women)* were asked to annotate a set of atomic sentences originating from the subsets of potential stereotypes or prevalent characteristics and their eight semantic variants as “abusive” or “not abusive”.

Who was involved in the data collection process?

Co-authors of the paper compiled the atomic sentences of the subsets “potential stereotypes” and “prevalent characteristics”, devised eight different semantic variants that each of the atomic sentences were converted to and also substituted the mention of the identity group by any of the other four groups. The annotation (“abusive” or “not abusive”) was carried out via crowdsourcing. Crowdworkers were recruited via the crowdsourcing platform *Prolific*[1]. They were compensated following the wage recommended by Prolific (i.e., $12.00 per hour). To ensure reasonable annotation quality, we chose the appropriate screening options and also permitted only crowdworkers with an approval rate of 100% and did not have dyslexia.

Over what time-frame was the data collected?

The data was collected during the second half of 2023 and the first half of 2024.

How was the data associated with each instance acquired?

The data was mostly observable as raw text, except that the labels were reported by subjects (e.g., survey responses). The potential stereotypes were obtained by browsing the Web and also asking members of these identity groups (i.e., recruited crowdworkers of Prolific) to provide a list. We also browsed the Web and collected some non-abusive sentences, prevalent characteristics, with a structural similarity to the potential stereotypes. In addition, we devised 8 different semantic variants of each atomic sentence of the two subsets. To obtain the data for the switched identity groups, we used the atomic sentences and substituted the mention of the identity group by any of the other 4 groups. The associated labels (i.e., abusive, not abusive, unknown) were obtained by asking recruited crowdworkers from Prolific.

Does the dataset contain all possible instances?

The dataset is a sample of instances.

If the dataset is a sample, then what is the population?

Samples were collected following the filtering steps that were applied at various stages during the process of data collection. Sentences that were considered not proper English, sentences that were explicitly abusive. We also excluded sentences being duplicates or near-duplicates to the sentences already included in the pool of collected sentences during this iterative creation process.

Is there information missing from the dataset and why?

No data is missing.

[1] https://www.prolific.co/

Are there any known errors, sources of noise, or redundancies in the data?

Since we applied filtering steps extensively, we hope to have reduced the level of noise in the resulting dataset to a minimum. As part of these filtering steps, duplicates and near-duplicates have also been removed from the final dataset. Therefore, we expect the dataset to contain no significant redundancies.

Data Preprocessing

What preprocessing/cleaning was done?

After obtaining the data (i.e., potential stereotypes and prevalent characteristics) by browsing the Web and also asking members of these identity groups (i.e., recruited crowdworkers of Prolific), we removed every sentence that contained obvious negative sentiment, which means negative polar expressions and negated positive polar expressions. The remaining sentences were normalized into atomic sentences with a simple, basic sentence structure. Based on the atomic sentences of both subsets, we devised 8 different semantic variants that each of the atomic sentences were converted to. By substituting the mention of the identity groups by any of the other 4 groups, we omitted any sentences where substituting identity groups resulted in clearly nonsensical outcomes.

Was the "raw" data saved in addition to the preprocessed/cleaned data?

Yes, but it is not part of the final dataset to be released publicly.

Is the preprocessing software available?

The above preprocessing had to be carried out manually. Therefore, no software is available.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

According to our extensive evaluation, we could show that all classifiers that have simply been trained on an existing dataset are unable to produce reasonable classification performance on our novel dataset. This can be interpreted as a proof that the specific type of abusive language that our novel dataset contains is not sufficiently represented in existing datasets.

Dataset Distribution

How is the dataset distributed?

The dataset is to be distributed via the first author's github account.

When will the dataset be released/first distributed?

It will be released upon publication of the research paper introducing this dataset "Beyond *Negative* Stereotypes – Non-Negative Abusive Utterances about Identity Groups and Their Semantic Variants".

What license (if any) is it distributed under? Are there any copyrights on the data?

The dataset is to be licensed under CC-BY-4.0. It will be made publicly available. There will be a request to cite the corresponding paper if the dataset is used: "Beyond *Negative* Stereotypes – Non-Negative Abusive Utterances about Identity Groups and Their Semantic Variants".

Are there any fees or access/export restrictions?

The dataset should be used for non-commercial purposes only, e.g. research.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The dataset is distributed via the first author's github account.

Will the dataset be updated?

No

If the dataset becomes obsolete how will this be communicated?

We do not foresee a scenario by which this dataset would become obsolete.

Is there a repository to link to any/all papers/systems that use this dataset?

A repository on github will be created allowing public access to this dataset.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

Others may do so and should contact the authors about incorporating fixes/extensions.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?

All crowdworkers contributing to the dataset were informed that the task they participated in was part of linguistic research.

If it relates to other ethically protected subjects, have appropriate obligations been met?

For the creation of this dataset, crowdworkers had to annotate potentially offensive language. Therefore, a respective warning in the task advertisement of the annotation task was included. The task description also stated that the researchers of this task pursue a linguistic purpose with these crowdsourcing tasks and that the opinion expressed in the sentences to be processed in no way reflects the opinion of these researchers.

If it relates to people, were there any ethical review applications/reviews/approvals?

Due to the delicate nature of the dataset, the legal department of the research facility at which this research was carried out was informed.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?

See answer to first question of this subsection.

If it relates to people, could this dataset expose people to harm or legal action?

Our dataset contains a subtype of abusive language that does not address specific individuals nor single persons. Since non-negative stereotypes also represent a fairly mild form of abusive language, we do not anticipate that this dataset could expose people to harm or legal action.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

It is a controlled dataset. The instances of non-negative stereotypes in our dataset encompass five specific identity groups (i.e., Black people, gay people, Jews, Muslims, women) but as only the recruited crowdworkers belonging to the respective identity group annotated the instances, we do not think that the dataset could be used to unfairly advantage or disadvantage a particular social group.

If it relates to people, were they provided with privacy guarantees?

Our dataset comprises non-negative stereotypes directed at five identity groups. All instances of abuse on that dataset do not target specific individuals. Therefore, such privacy guarantees are not applicable as far as the targets of non-negative stereotypes is concerned.

The crowdsourcing platform we use, i.e. Prolific, does not provide the identity of the crowdworkers participating in a particular task. Therefore, we consider the privacy of the crowdworkers to be guaranteed.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)?

We have no indication that our dataset is an any way non-compliant with GDPR.

Does the dataset contain information that might be considered sensitive or confidential?

Since our dataset does not target specific individuals and the crowdsourcing platform we use, i.e. Prolific, does not provide the identity of the crowdworkers, we think that this is not the case.

Does the dataset contain information that might be considered inappropriate or offensive?

Due to the nature of the research task addressed, the dataset contains a significant amount of offensive language.