This datasheet refers to the dataset comprising ***sentences collected from Twitter for the task of implicitly abusive language detection with identity groups representing the abused target****. It refers to both the English and the German version of the dataset.*

Motivation for Dataset Creation

Why was the dataset created?

The dataset was created to enable research on identifying implicitly abusive remarks about identity groups. Previous datasets for abusive language detection are extremely biased when it comes to abusive remarks on identity groups. In our focused dataset, we applied several measures to ensure that both abusive and non-abusive sentences referring to identity groups are equally represented. Our dataset also focuses on atomic utterances.

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

This is a focused dataset for a subtask of abusive language detection only. Beyond the experiments we carried out in this paper, this dataset could be used as a further functional test for HateCheck[1] since the phenomena we address with our new dataset are not represented in HateCheck.

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?

Each instance represents an atomic sentence extracted from Twitter in which an identity group is mentioned as the agent of the main verb. The main verb is a negative polar verb. Thus, we want to ensure that the sentence conveys an overall negative polarity. We chose four identity groups: homosexuals[2], Jews, Muslims and women.

Are relationships between instances made explicit in the data?

[1] https://github.com/paul-rottger/hatecheck-data

[2] To be precise, we used the two expression *gay people* and *lesbians*, since the term *homosexual* already has an (implicitly) offensive connotation (https://www.psychologytoday.com/us/blog/speaking-in-tongues/202105/why-is-the-word-homosexual-considered-be-offensive)

No. The dataset was sampled from a wide set of different users in order to avoid a user bias. On average there are only 2 instances from the same user. In general, our dataset does not provide any information on the users. Therefore, this dataset is hardly suitable for studying the relationships between instances.

How many instances of each type are there?

The English version of the dataset comprises 2221 instances, whereas the German version comprises 970 instances. *(The German dataset is smaller since there are much fewer tweets available in German than in English. The general uneven number of instances in both versions of the dataset are the result of applying many filtering steps in order to obtain unbiased data.)*

What data does each instance consist of?

For each instance, we provide the following information:

- The sentence itself.
- The binary label indicating whether the instance was rated as abusive or not.*
- The (negative) main verb in lemmatized form.
- The target, i.e. the mention of the identity group. This is always the agent of the main verb.
- The binary label specifying the aspect type: It indicates whether the sentence is either episodic or non-episodic.*
- The binary label indicating whether the agent is depicted as a perpetrator or not.*
- The patient of the main verb. (We provide the entire noun phrase.)
- The binary label indicating a priori sentiment of the patient. We distinguish between positive and negative sentiment. (We conflated neutral and positive sentiment since it did not make a difference with regard to our experiments.)*
- The binary label indicating the sentiment of the agent towards the patient as evoked by the main verb. Again, we distinguish between positive and negative sentiment.*
- The binary label indicating whether the agent (i.e. the identity group) is attributed to a non-conformist view. There is such attribution if the sentiment indicating the (a priori) sentiment of the patient and the sentiment of the agent towards the patient do not agree.

*These labels have been established via crowdsourcing, i.e. they are the result of manual annotation. Each label represents the majority label over ratings provided by 5 different crowdworkers.

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

Everything is included.

Are there recommended data splits or evaluation measures?

Both the English and the German version of the dataset are thought to be used as test sets only. In our experiments, we always trained on another dataset in order to avoid overfitting to dataset artefacts. Therefore, no splits are provided.

What experiments were initially run on this dataset?

A linguistically-informed classifier and generic supervised classifiers (i.e. transformers) which were trained on other datasets were tested on this dataset. The task was the binary classification to determine whether a sentence comprises abusive language or not.

The linguistically-informed classifier comprises three component tasks:

- one component classifier that distinguishes between episodic and non-episodic sentences
- one component classifier that determines whether the agent (i.e. the mention of the identity group) is depicted as perpetrator or not
- one component classifier that determines whether the agent (i.e. the mention of the identity group) is attributed to a non-conformist view or not

For the linguistically-informed classifier, we tested an automatic classifier and an oracle-version in which gold labels for the component tasks were combined instead of using actual classifiers.

Data Collection Process

How was the data collected?

Tweets from Twitter were collected by searching for mentions of identity groups followed by a negative polar verb.

Filtering steps were applied in order to obtain a fairly unbiased dataset:

- Our data is sampled from one textual source. Both abusive and non-abusive sentences are sampled by the same pattern (see above). Thus, no biases are caused by merging instances from different text sources.
- In order to avoid any user biases, tweets were sampled from a wide set of different users. The average amount of tweets per user is about 1.1.
- In order to avoid a focus on frequently occurring verbs, the dataset was sampled from a large set of negative polar verbs. On average each verb occurs twice in the final dataset.
- Only sentences that do not include any explicitly abusive words, were included. Otherwise, classifiers could easily detect the respective abusive utterances since they would just have to focus on these explicit clues.
- All texts co-occurring with our sentences that might give rise to spurious correlations, e.g. hashtags or user names, were removed.

Who was involved in the data collection process?

Co-authors of the paper were involved. The manual annotation was done via the crowdsourcing platform *Prolific*[3]. All crowdworkers were compensated following the wage recommended by Prolific (i.e. $9.60 per hour).

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

Over what time-frame was the data collected?

The data on homosexuals, Jews and Muslims was collected during 2020 and the first half of 2021. The data on women was collected in fall 2021. The collected data represents tweets that were available at the time of collection. They were not restricted to the tweets posted during that period. (Focusing on a restricted time frame would have meant that the resulting dataset would have become too small.)

How was the data associated with each instance acquired?

The data was observable as raw text.

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 as outlined for the question “How was the data collected?”. All instances that remained after applying the filtering steps were considered for manual annotation (via crowdsourcing). A few instances were removed afterwards, if the crowdworkers deemed the quality of the language of a sentence insufficient (e.g. if they could not understand the content of the sentence).

Is there information missing from the dataset and why?

No data is missing.

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

Due to the nature of Twitter, the quality of language may vary. As long as crowdworkers did not consider the language of a sentence too poor (e.g. so that one could not understand its content), the instances were kept in the final dataset. Therefore, orthographic or grammatical errors may occur in the sentences of the dataset. Duplicates have been removed from the final dataset.

Data Preprocessing

What preprocessing/cleaning was done?

Preprocessing was only necessary for running the linguistically-informed classifier on the dataset (e.g. part-of-speech tagging and lemmatization). For generic supervised classifiers (i.e. transformers) no preprocessing was necessary.

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

Not applicable. The "raw" data represents the actual dataset.

Is the preprocessing software available?

No

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

This is the first dataset to enable focused research on negative (atomic) sentences involving identity groups with regard to abusive language detection. Heavy filtering was applied in order to create the most unbiased dataset possible. Most classifiers trained on existing datasets for abusive language detection and tested on this novel dataset only produced poor performance scores. This indicates that this novel dataset may more adequately represent instances for this subtask in abusive language detection than previous datasets.

Four different identity groups are considered as (potential) abused targets in this dataset, i.e. homosexuals, Jews, Muslims and women. Therefore, this dataset enables the detection of predictive linguistic properties that can be observed across different identity groups.

This dataset is a focused dataset. It has not been devised as a general dataset for abusive language detection. It is only designed for enabling research on implicitly abusive language detection involving negative (atomic) sentences on identity groups.

Dataset Distribution

How is the dataset distributed?

The dataset is 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 "Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach".

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: "Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach".

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 original 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?

No. The data was extracted from public web sources. There was no explicit informing of these authors that their posts were to be used in this way. The public release of a limited number of tweets as in the range of our dataset is in accordance with the regulations of Twitter. Moreover, in order to protect the privacy rights of the authors of the sentences and individuals mentioned in them, we anonymized the dataset by discarding mentions of usernames.

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 annotated 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 first question of this subsection.

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

The data was already public. However, we anonymized the dataset by discarding mentions of usernames in order to prevent people being exposed to harm or legal action.

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

In many instances, the identity groups mentioned are subject of verbal abuse.

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

No. However, we anonymized the dataset by discarding mentions of usernames.

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

It may comply with that regulation since the dataset was anonymized.

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

We removed personally identifying information. However, the dataset may contain several instances of group-related incitement.

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 considerable amount of anti-Semitic, homophobic, Islamophobic and misogynist content.