Motivation for Dataset Creation

Why was the dataset created?

The dataset was created to enable research on euphemistic abuse, 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 euphemistic abuse, we also envisage its use for evaluating the *generation* of paraphrases, for example, those paraphrases representing explicitly abusive sentences that are derived from their euphemistic counterparts. This was also examined in brief in our paper by the sequence-to-sequence approach.

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 a sentence. It is either an instance of euphemistic abuse or some other (non-abusive) sentence. The latter instances have been chosen in such a way that they are similar to the sentences representing euphemistic abuse. (This was on purpose, since such negative data are known to be effective for building supervised classifiers.)

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 600) 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 1797 English sentences in total, where 640 sentences represent some form of euphemistic abuse and the remaining 1157 sentences are non-abusive sentences.

What data does each instance consist of?

For each instance, we provide the following information:

- The sentence itself.
- 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.
- The corresponding fragment. (Fragments represent substrings of the sentence. They were derived from euphemistic abusive sentences. As such, they should be neutral, that is, they should not convey abuse. The fragments were used to create non-abusive sentences that are similar to the sentences representing euphemistic abuse.)
- For each sentence we also include the manually extracted features. These are binary features (they are not mutually exclusive) produced by the co-authors of the paper that indicate the presence of the following high-level concepts:
    - negated antonyms of abusive words
    - opposing sentiment
    - taboo topics
    - "extreme" language
    - lexicalization of euphemistic abuse
    - attribution of unusual properties to the target
- For each sentence representing euphemistic abuse, we also stated the original cue phrase (i.e. a sentence with explicit abuse) which had been used as stimulus to the crowdworkers in order to produce a euphemistic paraphrase.

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 such a way that sentences of the test set exclusively comprise euphemistic abuse referring to cue phrases not observed in the 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 two forms of classifiers, cross-dataset classifiers, i.e. supervised classifiers that were trained on a different dataset/task, and within-dataset classifiers, i.e. supervised classifiers that were both trained and tested on the new dataset (via 5-fold cross-validation).

As cross-dataset classifiers, we considered a classifier trained on word definitions ("definition-based classifier"), a classifier to detect general forms of euphemisms and various classifiers trained on previous datasets for abusive language detection.

As within-dataset classifiers, we examined various transformers fine-tuned on the new dataset and a feature-based classifier that combines a small set of high-level features that were manually extracted.

Data Collection Process

How was the data collected?

The dataset was not collected from existing sources. Instead, crowdworkers were asked to invent euphemistic paraphrases given an explicitly abusive sentence (i.e. cue phrase). Non-abusive sentences were obtained by asking crowdworkers to invent a non-abusive sentence that includes a given (neutral) text fragment derived from a euphemistic abusive sentence.

Who was involved in the data collection process?

Co-authors of the paper compiled the cue phrases and the (neutral) fragments that were derived from the invented sentences representing euphemistic abuse. The actual sentences (both abusive and non-abusive) were produced via crowdsourcing. Crowdworkers were recruited via the crowdsourcing platform *Prolific*[1]. They were compensated following the wage recommended by Prolific (i.e. $9.60 per hour).

Over what time-frame was the data collected?

The data was collected during the second half of 2021 and the first quarter of 2022.

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 that were applied at various stages during the process of data collection. Sentences that were considered not proper English, sentences that were explicitly abusive, and abusive sentences that did not match the given cue phrase semantically were excluded. 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?

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

No data is missing.

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

Since we applied filtering steps extensively (*see our main paper for more details*), 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?

Some of the crowdworkers who invented sentences of our dataset displayed some idiosyncratic writing style, e.g. all words written lowercase or non-standard usage of capitalization. As each crowdworker only contributed to one of the two classes, we had to normalize these writing styles as otherwise, this could have led to artefacts in the resulting dataset: supervised classifiers could have benefitted from learning spurious correlations between such idiosyncratic writing styles and a particular class. By normalizing all sentences to a standard spelling such spurious correlations could not stand a chance to emerge.

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 not able 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 "Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language".

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

The dataset is to be licensed under CC BY-NC-SA 4.0. It will be made publicly available. There will be a request to cite the corresponding paper if the dataset is used: "Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language".

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 or even create it. 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 or groups of individuals. Since euphemistic abuse also represents 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?

The instances of euphemistic abuse in our dataset do not target social groups at all. Therefore, 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 euphemistic abuse directed at a generic second person. 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 euphemistic abuse 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 or groups of 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.