

📁

annotators

📁

data

📁

documentation

📁

results

📁

scripts

📁

semantic\_annotation

📄

README.md

📄

main.py

📄

readme\_preview.pdf

📄

requirements.txt

Last Update: Apr 26, 2024

Support me

README.md

Edit

View raw

Download file

python 3.12+

# HateRep: A Diversified Hate Speech Dataset with Repetitions

This dataset contains 2880 annotations from participants with diverse gender and sexual orientations. Social media posts have been extracted from for existing hate speech databases ([Measuring Hate Speech](#), [Gab Hate Corpus](#), [HateXplain](#), and [XtremeSpeech](#)).

The methodology and findings from this study are presented in: **Enhancing Hate Speech Annotations with Background Semantics**

## Repo structure

Data is organised in the following folders:

- Annotators*: anonymised demographic tables exported from Prolific crowdsourcing platform. Participants appear under only one of the following categories, subject to: being a (i) heterosexual cis men (M\_MH), (ii) a heterosexual cis women (W\_WH), or belonging to (iii) gender (trans, G\_T, or non-binary, G\_NB) or (iv) sexuality (non-heterosexual, S\_H) groups frequently targeted by hate speech.
- Data*: contains semantic and crowdsourcing annotations. The specific annotation categories are shown in the [figure](#) below.
- Semantic\_annotation*: contains the background knowledge of the hate speech sample, which was mainly provided by a domain-specific KG, i.e., the [GSSO](#) (`pruned_concepts.csv`) and completed with generic semantic resources (`missing_concepts.csv`).
- Documentation*: contains the approved Ethics Application Form and Participant Information Sheet.

Source code is in *scripts*, specifically in the Python files:

- dataCollect.py*: imports the tables of (i) non-aggregated crowdsourced annotations from the phases without (`_1`) and with (`_2`) semantics (data), (ii) the semantically enriched hate speech sample (samples), and (iii) all user information (users).
- agreement.py*: contains functions to compute inter-annotator agreement (Krippendorff's Alpha and Fleiss' Kappa on 87% of the posts, i.e., with 6 annotations).
- helper.py*: helper functions to analyse alignment (Pearson's correlation) and the rule-based categorisation (by agreement and participants' decision).
- utils.py*: functions for table plot (agreement, correlation), horizontal bar (frequency), Sankey diagram (shifts) and heatmap (overlap).

All files used for evaluation in the paper are in folder *results*.

## Hate Speech Annotation Example

Post:

the [kebabs](#) are a bunch of [homosexual rapist](#) deviants, so they 'll keep it alive as a [sex slave](#).

s://en.wiktionary.org/wiki/kebab

Highlighted definitions:

'homosexual'

'The state of being sexually and romantically attracted primarily or exclusively to persons of a gender identity the same as one's own.'

Part 1: Identify if there are any references to gender and/or sexuality.

Does the message mention or is about gender?

Select any or all that apply.

☐ Men ☐ Women ☐ Non-binary ☐ Other gender

☐ It specifically mentions or is about transgender.

Or select one option.

https://anonymous.4open.science/r/hateRep-A240/README.md

1/1