General Characteristics

1.1 Who collected the dataset, and who funded the process? The dataset was collected by a team of researchers associated with the GESIS - Leibniz Institute for the Social Sciences CSS department. In particular: Samory, Mattia (Head Author); Sen, Indira (Data Curator); Kohne, Julian (Data Curator); Floeck, Fabian (Project Lead); Wagner, Claudia (Project Lead).

1.2 Where is the dataset hosted? Is the dataset distributed under a copyright or license? The dataset is hosted at the GESIS datorium (DOI: https://doi.org/10.7802/2251), and can be freely accessed after registration. It is distributed under the CC BY-NC-SA 4.0 license.

1.3 What do the instances that comprise the dataset represent? What data does each instance consist of? The instances in the dataset represent scale items, tweets, or adversarial augmentations created by crowdworkers. The instances have the following data fields: a unique ID, the dataset they are coming from, a toxicity score from the Perspective API[1], a binary sexism classification, and a second ID if the instance is a modification of another instance. A supplementary dataset, providing the individual labels for sexism due to both content and phrasing, as well as the annotations from each crowdworker that, in aggregate, inform the final sexism classification, is also available.

1.4 How many instances are there in total in each category (as defined by the instances' label), and - if applicable - in each recommended data split? There are a total of 13,631 (1,809 labeled as sexist) instances in the dataset. Of these, 2,292 instances are adversarial examples. There are different sources for the instances, with the following numbers of instances coming from each of the sources:

- 1,080 (189) from the benevolent dataset, 402 of which are adversarial augmentations
- 2,431 (790) from the callme dataset, 1,151 of which are adversarial augmentations
- 1,257 (290) from the hostile dataset, 579 of which are adversarial augmentations
- 878 (540) from the scales dataset, 135 of which are adversarial augmentations
- 7,985 (0) from other datasets, 25 of which are adversarial augmentations

[1] https://www.perspectiveapi.com/

1.5 In which contexts and publications has the dataset been used already? The dataset has been used by Samory et al.(2021) to evaluate the reliability of machine learning models trained for the task of detecting sexism in tweets. It has also been used in follow-up work by Sen et al. (2022) for assessing the efficacy of adversarial augmentation.

1.6 Are there alternative datasets that could be used for the measurement of the same or similar constructs? Could they be a better fit? How do they differ? While there are many datasets available for different aspects of sexism on social media platforms, this dataset proposes a method to map out the various aspects of sexism as a construct as comprehensively as possible, allowing to operationalize and measure the construct of sexism as nuanced as possible. Depending on the nuance and dimension of sexism a researcher is interested in, any of these specific datasets might be an appropriate fit, while this dataset works as a more comprehensive and general one. Other alternative datasets could be related to abusive language in general as well as more specific datasets on misogyny.

1.7 Can the dataset collection be readily reproduced given the current data access, the general context and other potentially interfering developments? While the scales used for the dataset should still be available as part of scientific publications (subject to license agreements), the authors release the redacted tweet texts rather than the IDs for safeguarding data subjects' privacy, which means that the dataset can be directly reused for future work. However, to replicate the data collection from scratch, the tweet IDs would be required.

% tweets (both those reused from other datasets as well as those collected for this dataset) might not be recoverable anymore. tweets could have been deleted, either by the authors or by the platform, and would thus not be available for collection anymore. The same would be the case if the author of a tweet decided to change their profile settings to private. % The general methodology, however, should still be replicable, allowing for the creation of similar, but updated versions of the dataset. However, the free-of-charge academic access to the Twitter API has since been discontinued, forcing researchers interested in reproducing the data collection to either sign up for one of the available paid API tiers or find alternative ways of collecting data from the platform.

1.8 Were any ethics review processes conducted? There was no IRB or ethics review process conducted. The work does not fall under research with human subjects and care was taken to ensure that crowdworkers were compensated in accordance with fair pay guidelines and debriefed about the potentially harmful nature of the content they would be annotating.

1.9 Did any ethical considerations limit the dataset creation? For ethical reasons, the adversarial augmentation was limited to turning sexist instances into non-sexist instances, and not vice-versa.

1.10 Are there any potential risks for individuals using the data? Does the data contain any disturbing images or texts? Could the content evoke psychological distress? Due to the nature of the dataset, i.e., manifestations of explicit and nuanced types of sexism, it can be potentially upsetting for data stewards.