Construct Definition

Validity

2.1 For the measurement of what construct was the dataset created? The dataset was created to measure the construct of sexism in a more comprehensive manner compared to state-of-the-art research, which is heavily focused on overt forms of sexism, with the goal of improving sexism detection online. The dataset creators provide detailed information on the literature and scales covered by their construct definition in Samory et al. (2021). Based on the scales collected from the literature, four new categories on sexism are iteratively formed: behavioral expectations, stereotypes and comparisons, endorsements of inequality, denial of inequality and rejection of feminism. Additionally, the dataset creators not only consider the contents of a text as potentially sexist, but also its phrasing.

2.2 How is the construct operationalized? Can the dataset fully grasp the construct? If not, which dimensions are left out? Have there been any attempts to evaluate the validity of the construct operationalization? The construct is operationalized through a coding scheme which is then used by annotators hired for the task of determining whether an instance is sexist or not. Those crowdworkers follow the instructions laid out in a codebook, available from Samory et al. (2021). The coding scheme was validated by asking five crowdworkers to apply the coding scheme to the ground truth data. The resulting annotations' majority verdict (min. three out of five) corresponded with the ground truth label in 86% of the cases. Further information on the measures taken to ensure qualitatively good and valid annotations are reported in response to questions *5.3*. Even though the dataset creators made an attempt to condense all the collected aspects and dimensions of sexism into their final categories, it is unlikely that textual content on social media covers all possible manifestations of sexism. As an example, sexism reflected in patronizing behavior towards women cannot be covered in this dataset.

2.3 What related constructs could (not) be measured through the dataset? What should be considered when measuring other constructs with the dataset? Scales referring to constructs "similar to sexism" are also included in the dataset. This could lead to issues related to convergent validity, if aspects outside of a specific definition of sexism are included in the dataset. However, for the same reason, the dataset could also be applicable to constructs closely related to sexism.

2.4 What is the target population? The authors do not explicitly define a target population for their study, but mention that their codebook should apply to "sexism in social media". The target population arising from this would thus be the population of all users of social media, or, if considering that the data collection is limited to English tweets, the population of English-speaking Twitter users.

2.5 How does the dataset handle subpopulations? The dataset does not explicitly include humans as subjects nor include demographic information of the creators of the different instances. It therefore does not explicitly identify, define, or act upon subpopulations.