Data Preprocessing and Data Augmentation

Trace Augmentation and Trace Measurement Error

5.1 Is there a label or target associated with each instance? If so, how were the labels or targets generated? Crowdworkers hired from Amazon Mechanical Turk were given the task of annotating the different instances for sexism. Each annotation consisted of one statement that crowdworkers were asked to annotate on two single-choice lists, containing the codes for sexist content and phrasing. The toxicity scores associated with each instance were obtained from the automated toxicity classifier of the Perspective API.

5.2 If automated methods were used, how does method performance impact the augmentations? The toxicity scores obtained from the Perspective API were derived automatically. The Perspective API runs on a ML model trained on comments from different online sources. This paradigm comes with a number of potential errors and biases that it might introduce through its annotations, as for example explained in this paragraph taken from the Perspective API's model card:
"Machine learning models learn from the data they're trained with, so any biases in the data can creep into the predictions the models make. For example, our models sometimes predict higher toxicity scores for comments with terms for more frequently targeted groups (e.g. words like""black", "muslim", "feminist", "woman", or "gay") because comments about those groups are over-represented in abusive and toxic comments in the training data."

5.3 If human annotations were used, who were the annotators that created the labels? How were they recruited or chosen? How were they instructed? How were they remunerated? Crowdworkers were recruited from Mechanical Turk and had to be located in the US, with over 10,000 HITs approved and over 99% HIT approval rate. They furthermore had to pass a strict qualification test that ensured that they understood the construct of sexism as defined in the codebook, and did not apply overly subjective notions of sexism in their labeling. The test was passed if they correctly annotated 4 out of 5 ground-truth samples. This final test could only be taken once. The training imposed on annotators before taking the test described the codes for each of the two annotation categories, relying on examples and counter-examples. As additional guidance, the codebook developed based on the sexism scales was available to the annotators. Crowdworkers received 6 cents per annotation, and 20 cents per counterfactual augmentation, resulting in a "fair hourly wage".

5.4 If the final label was derived from multiple annotations, how was this done? Each instance was annotated by five annotators, and only those instances which at least three annotators considered as sexist, either because of content or phrasing, were marked as such. The final label was then obtained via majority vote.

5.5 Have there been any attempts to validate the labels? For the annotations generated by the crowdworkers, and using the majority rule for the five annotations provided for every sentence, Randolph's Kappa for the content-annotations ($\kappa = .62$) and phrasing-annotations ($\kappa = .82$) were calculated and found to be satisfactory. The authors additionally report a majority agreement of 81%, 98.8% and 100% for content, phrasing and overall sexism, respectively.

5.6 How could the data be misused? As the data contains texts that were identified as sexist and/or toxic, together with corresponding labels or scores, potential misuse scenarios of this data would include their utilization as a source for sexist content - be it by directly "re-using" the instances of the dataset, or by using them as input or training data for systems that generate sexist content.

5.7 Could the dataset in any way contribute to the creation or reinforcement of social inequality? The dataset could potentially be used - together with other datasets of a similar type - to train or finetune a generative language model for the creation of sexist tweets, similar to those included in the dataset. Such a model could then be used generate and populate social media and web platforms with sexist content to harass vulnerable people, thereby reproducing and amplifying the harms and inequalities caused by the use of sexist language online.

User Augmentation Error

5.8 Have attributes and characteristics of individuals been inferred? There are no individuals explicitly identified in the dataset, thus there are neither observed nor inferred attributes and characteristics of individuals in the dataset.

5.9 Is it possible to identify individuals either directly or indirectly from the data? As there are no explicitly identified in the dataset, and as those individuals being referred to or mentioned in tweets have been pseudonymized, it should not be possible to neither directly nor indirectly identify individuals from the data.

Trace Reduction Error

5.10 Have traces been excluded? Why and by what criteria? No, all tweets collected via the previously described query have been retained.

User Reduction Error

5.11 Have users been excluded? Why and by what criteria? No, all Users corresponding to the tweets collected via the previously described query have been retained.

Adjustment Error

5.12 Does the dataset provide information to adjust the results to a target population? If so, is this information inferred or self-reported? The dataset does not contain any demographic

information, and thus does not contain any information that would allow for any type of inference.