Platform Selection

Platform Affordances Error

3.1 What are the key characteristics (relevant to the collected data) of the platform at the time of data collection? Key characteristics of Twitter at the time of data collection include the 280 character limit for tweets.

3.2 What are the effects of the terms of service of the platform on the collected data? As the terms of service (ToS) at the time of data collection have not been documented, the following is based on the current version. The ToS state that the user is responsible for any content they provide, as well as for their compliance with applicable laws, rules, and regulations. It furthermore identifies severe, repetitive usage of [...] sexist tropes where the primary intent is to harass or intimidate others as a reason for a tweet's removal. Through these rules and their enforcement through moderation practices, Twitter tries to remove sexist content from the platform. Therefore, the efficiency and efficacy of the moderation practices influences how many sexist tweets are available for collection from the platform at any point in time, and the design of the moderation practices determines what types of sexist content are removed and what types remain on the platform. Changes in the ToS and changes in their enforcement through moderation practices would thus be reflected in the collected data.

3.3 What are the effects of the sociocultural norms of the platform on the collected data? At the time of data collection, Twitter was well established as the most popular platform for certain types of users to comment on acute events and issues of general importance for the society. The debates on Twitter are often perceived as polarized and certain topics tend to "blow-up" on Twitter, being disputed about at great length and with great fervor by users. This has led to instances of trolling, where users would try to trigger such (exaggerated) reactions from other users by posting provocative, polarizing, or even straight-out abusive and harmful statements, including sexist tropes, even if with an implied distance or irony to it, or in some sort of meme-format, as with the phrase "Call me sexist, but.." . The dataset is directly picking up on these patterns that are the result of the culture on Twitter.

3.4 How were the relevant traces collected from the platform? Are there any technical constraints to the data collection method? If so, how did those limit the dataset design? The tweets were collected via the Twitter streaming endpoint of the Twitter API. Tweets with the phrase "call me sexist, but", posted between 2008 and 2019, were collected from the API. Collecting data

from the historic search endpoint means that, in theory, every tweet since the creation of Twitter back in March 2006 can be searched and thus potentially be included in the dataset. However, retrospective collection of tweets, as is always the case when using the historic search endpoint, does not allow for the collection of tweets that have been removed by Twitter, by the user, or that have been made private. Twitter might remove tweets that directly violate their rules, and also automatically removes tweets that were posted by users that have been (permanently) removed from the platform, e.g. for their failure to comply with the rules. A tweet would also not be accessible anymore if its author decided to delete it entirely, or if that author changed their account settings to private, thus preventing the tweet from being found by others.
For this dataset, this could mean that not all tweets that were ever posted containing the phrase call me sexist, but are included in the dataset. Since the phrase is chosen for being and indicator of potentially sexist content that follows, it is rather likely that some of the tweets containing that phrase have either been deleted by the user for the backlash that the tweet received, or for further reflections that made the author uncomfortable with being associated with the tweet's contents. The sexist nature of these tweets also makes it more likely that Twitter removes them due to content moderation, either by directly removing the tweet or indirectly, by suspending its author from the platform.

3.5 In case multiple data sources were used, what errors might occur through their merger or combination? The final dataset is a combination of different existing datasets for sexism on social media, a collection of psychological scale items used to measure sexism, as well as a newly collected Twitter dataset. Challenges of combining these different data sources include the mixture of different types of texts (i.e. scale items and tweets), which very likely have different characteristics, e.g. in terms of their length, their structure, their linguistic style, or their vocabulary. To a certain extent, this also holds true for the different datasets collected from Twitter, where differences in the collection strategy, the time of collection, or the pre-processing of the raw tweets might lead to artifacts in the data. However, by using a codebook specifically derived for this dataset and by re-annotating the entire dataset, a lot of effort is made to ensure the compatibility of instances coming from different sources.

Platform Coverage Error

3.6 What is known about the platform population? The Pew Research Center regularly conducts surveys to characterize the Twitter platform population (Wojcik and Hughes, 2019). They do so by asking their survey participants on their Twitter usage, as well as by asking for consent to analyze their Twitter data. Their findings are then reweighted to match the target population, which is Twitter users age 18 and older, living in the U.S.. For these adult Twitter users, it is found that media personalities, politicians and the public turn to social networks for real-time information and reactions to the day's events. Twitter users, however, tend to be younger, are more likely to identify as Democrats, are better educated and have higher incomes than the overall population of U.S. adults. In themselves, Twitter users split into those that are very active, with the 10% most active users being responsible for 80% of all tweets in the U.S., and those that are inactive or turn to the platform only to keep up to date and to

read, not to actively participate and engage. While there are many anecdotal insights into the demographics and motivations of Twitter users, only few actually reliable empirical analyses of the platform population exists, as many Twitter users use the platform anonymously or pseudonymously, or simply do not provide the relevant demographic information. Researchers then need to resort to inferring those missing demographics, which in many cases might be of questionable reliability (Buolamwini and Gebru, 2018).