Perception of Offensiveness in Social Media:

The Influence of Annotator Demographics on Targeted Content

1. Abstract

How do annotators' demographics influence their perception of offensiveness in social media posts? And how does this effect change for posts targeted towards gender?

To answer these questions, we use the offensiveness rating annotations from the POPQUORN dataset, the Social Bias Inference Corpus (SBIC), and a merged version of the data. Firstly, we fit a linear mixed effects model with offensiveness as the dependent variable and the demographics of the annotators as independent variables. We find inconsistent results across the datasets for gender and race. For age, we find that annotators older than 60 tend to rate posts as more offensive than younger annotators. To answer the second question, we use a dictionary approach to detect posts targeted towards gender. We then examine the effect of annotators' demographics on their perceptions of offensiveness using subsets of the data. One contains only posts targeted towards gender and the other contains the remaining posts. Again, the results are not consistent across the data. In the SBIC data, women seem to annotate posts as even less offensive when the content of the post is targeted at gender. This is not what we expected.

Finally, we validate the dictionary approach using target category annotations from the SBIC and find that the approach does not perform very well. In addition, we find that the unexpected results we found for our second question could be caused by our dictionary method.

2. Motivation

Social media has a huge impact on many people's lives. It is important for users and the platform that the platform is a safe space where people do not feel harassed and can trust the information they consume. This is why content regulation plays an important role. Today, as the scale of these platforms has grown enormously (Gillespie,2020), AI models are a helpful tool for regulating content on the platforms. For training these models, annotations are crucial, for example in the case of hate speech and toxicity detection. However, what if identifying offensive content is not so straightforward, and the demographics of the annotators or the target of the offensive content would systematically influence how workers annotate? If this was the case, we could end up with biased models that under- or overestimate the offensiveness of posts. Therefore, when training the models, it would be important to have a balanced sample in terms of demographics to make the model as representative as it can be, or to correct for such bias through model parameters. This is also a big challenge because the language used on these platforms changes rapidly, and the perception of offensiveness is subjective and can change over time. It would therefore be very important to retrain and update these models on a regular basis, as well as to examine the effect of annotator demographics from time to time.

3. Research background

In prior research, sociodemographic backgrounds have been identified as crucial in shaping annotators' perceptions in subjective NLP tasks, leading to efforts to integrate such information into predictive models (Luo et al., 2020; Sap et al., 2022; Kumar et al., 2021). While these studies highlight the importance of sociodemographic, they also reveal challenges, including the potential amplification of biases when models are prompted with this data (Blodgett et al., 2020; Santurkar et al., 2023).

Pei and Jurgens (2023) use the POPQUORN Dataset to study the influence of annotator demographics on their perception of offensiveness. While they do not find statistically significant difference between men and women, they find that people older than 60 tend to find posts as more offensive than middle-aged participants. They also find notable difference in how people from different racial groups perceive offensiveness with Black participants rating the same content as significantly more offensive implying the risk of Bias in annotated Data. Theres also evidence, that while sociodemographic attributes might not always enhance model performance in annotation tasks, they are still key factors to consider (Orlikowskiet al., 2023). The results suggest that while sociodemographic background may not add additional learning beyond inherent variation, this does not negate their usefulness. Instead, it highlights the importance of using sociodemographic data judiciously, as its effectiveness may vary depending on the task and model architecture, suggesting the need to carefully study the influence of annotators demographics on their perceptions, especially when targeting specific groups.

The underlying theory motivating our research topic is the Social Identity Theory (Tajfel and Turner, 1979) suggesting that individuals tend to favour their in-group and are more protective of it., perceiving content as more offensive and harmful when targeting their own in group. Related work shows that people belonging to a specific sociodemographic group (e.g., LGBTQ people) rate content as more toxic when it targets their own in group (Goyal al., 2022) motivating our research to study the influence of gender on posts that specifically target gender.

4. Data

To answer our research questions, we use the offensiveness rating annotations from the POPQUORN Dataset and the Social Bias Inference Corpus (SBIC). Both Datasets contain posts with their respective offensiveness annotation and demographic characteristics of the annotator. The POPQUORN Dataset is designed to study how social media users annotate and perceive the offensiveness of online content using annotations from Reddit comments, whereas the SBIC is designed to study and model social biases present in online text using annotated posts from Reddit, Twitter, and Hate Sites. The SBIC provides us with three annotations per post, whereas in the POPQOURN Dataset, each comment is annotated approximately nine times. While in the SBIC offensiveness is measured as a categorical variable *(1: yes, 0.5: maybe, 0: no),* in the POPQUORN Dataset it is measured using a 1-5 Likert scale *(1 == "not offensive at all, 5== "very offensive")*.

To understand the composition of our datasets and how it might influence the perception of offensiveness, we analyse the distribution of gender within each race group in both Datasets. across different race groups for the POPQUORN dataset and the SBIC. We use a logarithmic scale to visualize the distribution of annotators, particularly for groups that are underrepresented. The distribution of both the POPQUORN Dataset and the SBIC show that most annotators are either

men or women, with non-binary annotators being significantly underrepresented. Furthermore, most of the annotators are white and outnumbering the other racial groups.

Figure 1 displays the log distribution of gender across different race groups for the POPQUORN dataset and the SBIC. We use a logarithmic scale to visualize the distribution of annotators, particularly for groups that are underrepresented. The distribution of both the POPQUORN Dataset and the SBIC show that most annotators are either men or women, with non-binary annotators being significantly underrepresented. Furthermore, most of the annotators are white and outnumbering the other racial groups. In the SBIC plot, the white and Asian groups have the highest number of annotators, with both genders (men and women) represented relatively equally. In contrast, the POPQUORN plot shows that the White group also has the highest number of annotators, but there is a noticeable decrease in the number of annotators across other racial groups, particularly for Black or African American, Asian, Native American, and Hispanic or Latino.

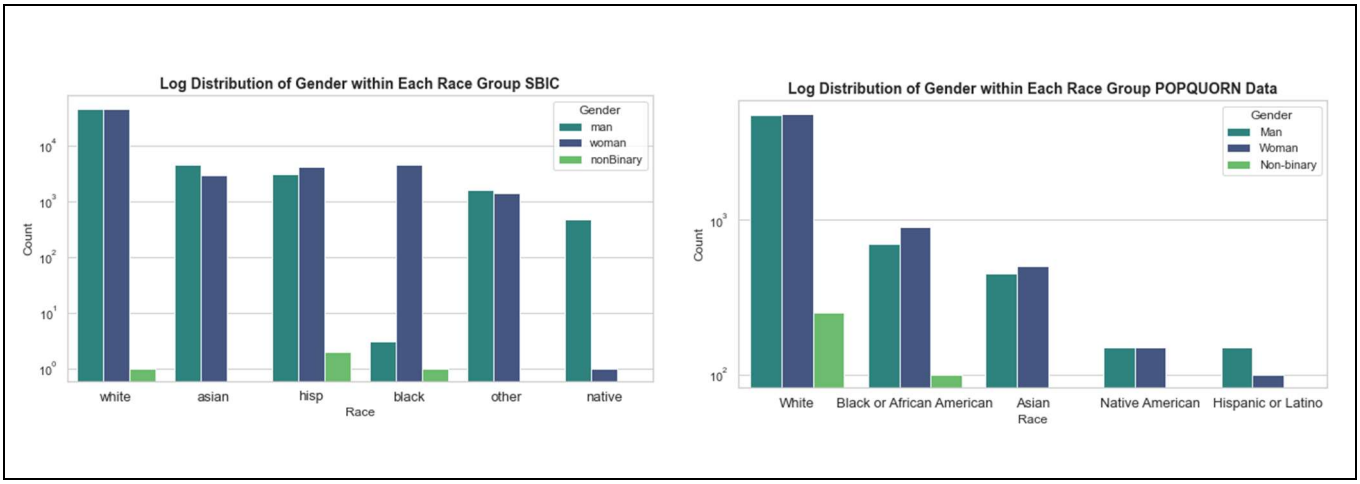

Figure 1: Log Distribution of Gender within Each Race Group

Average Offensiveness

When taking a first look into the mean offensiveness ratings of the gender groups in each dataset we can observe that men tend to rate posts as slightly more offensive than women and non-binary persons in the POPQUORN Dataset (Figure 2[1]). In the SBIC non-binary persons have an average offensiveness score of one, which is the highest rating, whereas men and women seem to rate offensiveness lower (~0.4).

Merged Dataset

To provide an enhanced dataset size and diversity we merge both datasets mapping the offensiveness values (ratings of 1.0 and 2.0 were mapped to 0.0, indicating low or no offensiveness; a rating of 3.0 was mapped to 0.5, representing moderate offensiveness, ratings of 4.0 and 5.0 were mapped to 1.0, indicating high offensiveness). The merged dataset contains 126.360 posts, with 12.987 posts from the POPQUORN Dataset and 142.938 from the SBIC. The distribution of the different racial groups (Figure 3[2]), with information about the dataset source additionally indicate the imbalance of the datasets as well as the imbalance of racial backgrounds in both datasets.

[1] Average Offensiveness Scores can be found in Appendix A
[2] Log Distribution of Source Dataset within Each Race Group can be found in Appendix B

5. Methods

First, analogously Pei and Jurgens (2023), we use a linear mixed effects model on the POPQUORN dataset and check if their findings can be replicated. Additionally, we see if they hold on to the Social Bias Inference Corpus and a merged version of the data. To answer our second research question, we classify the target group of each post, if any, into categories using a dictionary approach. We separate the posts that target gender from those that do not. Finally, we run the linear mixed effects regression on the subsets of the data. Both methods, the linear mixed effects model and the dictionary approach, are explained in the following section.

5.1. Linear Mixed effects model

Based on Pei and Jurgens (2023), offensiveness is modelled using a linear mixed effects model. We include the annotator's gender, age, and race as predictors for the offensiveness and control for post-specific effects. In this way, only deviations from the mean offensiveness of each post (here referred as instance) are analysed. Formally, the offensiveness of instance i rated by person j can be modelled as

$$Offensiveness_{ij} = \beta_0 + Z\,\alpha_i + \boldsymbol{\beta X_{ij}} + u_{ij} \qquad (1)$$

Here $Z\alpha_i$ represents the random effects associated with each instance i. Note that they are assumed to have zero mean and finite variances, and to be uncorrelated. $\boldsymbol{\beta}$ is a vector that models the effects of the demographics of the annotators, which in our case are mostly categorical variables. β_0 is the intercept, here associated with the reference category of our categorical variables and u_{ij} is the random deviation of each annotation.

5.2. Dictionary method for target identification

To assign instances to different target categories, we use the dictionary created by Yu et al. (2024). They built on related work that uses keywords from hatebase.org, then they identified the terms that were most indicative of hate speech, removed context sensitive words and created new categories. The final dictionary contains the target categories: age, body, class, disability, gender, nationality, organisation/institution, political, race, religion, and sexuality.

For the dictionary method to be effective, we perform some minor pre-processing steps on each of the instances: Delete stop words, digits, punctuation, convert the text to lower case and singularise nouns. Since some of the terms in the dictionary are bigrams, we not only create unigrams, but also bigrams from each instance. We then classify an instance into a target category if one of the unigrams or bigrams matches the terms in the dictionary. Note that an instance can be in several categories. The final distribution of the different datasets across targets is shown in appendix C. Figure 4 shows that, there are roughly twice as many untargeted instances as targeted instances, and gender is the most common target category among the targeted posts. For our analysis, we will only focus on posts that target gender versus those that do not, so we modify the categories to *target is gender* and *target is not gender*. Note that the category *target is not gender* also includes untargeted posts.

6. Results

6.1. RQ 1: How do annotators' demographics influence their perception of offensiveness in social media posts?

The first column of Table 1 shows the results of the Linear Mixed Effects model from equation (1) on the POPQUORN dataset. We run this regression to see if we can replicate the results of Pei and Jurgens (2023). They do not find a significant effect of gender and find that black annotators and annotators over the age of 60 tend to rate post as more offensive than the other groups within the variable. Unlike them, we do not include education as it was not available for the SBIC. Still, our results are very similar to theirs. For gender, both female and non-binary have negative coefficients. However, the coefficient for female is not significant. This seems to indicate that non-binary annotators might be more likely to rate the content as less offensive. For race all coefficients are significant. Black is the only coefficient with a positive sign, meaning that black annotators tend to rate content as more offensive, followed by Asian annotators. For age, annotators over 60 perceived higher offensiveness scores than younger annotators.

Table 1: LME for different datasets

	POPQUORN	SBIC	merged
Intercept	1.995***	0.464***	0.124***
	(0.046)	(0.005)	(0.011)
Dataset[T.SBIC]			0.312***
			(0.010)
gender[T.Woman]	-0.018	-0.021***	-0.014***
	(0.019)	(0.002)	(0.002)
gender[T.Non-binary]	-0.225***	0.085	-0.071***
	(0.060)	(0.163)	(0.016)
race[T.Black]	0.173***	0.064***	0.078***
	(0.044)	(0.006)	(0.006)
race[T.Hispanic or Latino]	-0.419***	0.128***	0.092***
	(0.077)	(0.005)	(0.005)
race[T.Native American]	-0.519***	0.045***	0.008
	(0.071)	(0.016)	(0.012)
race[T.White]	-0.113***	0.082***	0.055***
	(0.036)	(0.004)	(0.004)
race[T.Other]		0.103***	0.080***
		(0.007)	(0.007)
age		-0.001***	
		(0.000)	
age[T.25-29]	-0.165***		0.042***
	(0.040)		(0.004)
age[T.30-34]	-0.166***		-0.006*
	(0.039)		(0.004)
age[T.35-39]	-0.158***		-0.002
	(0.038)		(0.004)
age[T.40-44]	-0.020		0.072***
	(0.042)		(0.005)
age[T.45-49]	-0.072*		-0.042***
	(0.043)		(0.005)
age[T.50-54]	-0.072		0.020***
	(0.044)		(0.007)
age[T.54-59]	0.017		0.025***
	(0.037)		(0.006)
age[T.60-64]	0.343***		0.137***
	(0.048)		(0.011)
age[T.>65]	0.148***		0.075***
	(0.039)		(0.008)
Group Var	0.424***	2.129***	1.932***
	(0.021)	(0.022)	(0.020)
N	12987	113373	126360

Standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01
Reference categories: Dataset [POPQUORN], gender[T.Men], race[T.Asian], age[T.18-24]

The second column shows the results of running the regression on the SBIC data before mapping. In this case, age is a numerical variable and offensiveness can only be 0(not offensive) ,0.5 (unsure) or 1(offensive). The results are different from what we found on the POPQUORN dataset: For gender, there is a significant negative coefficient for women. For race, Asian annotators seem to rate content as less offensive than all other categories, while in POPQUORN, they only rate content as less offensive than black annotators. Age has a negative coefficient, meaning that older annotators would rate content as less offensive. This could be because in this dataset age is not a categorical variable.

As the SBIC is over-represented in the merged version of the data, the results from the merged dataset are similar to those from the SBIC. Men find the posts more offensive than women, and non-binary and Asian annotators seem to find the posts more offensive. Similar to Popquorn, we find that older groups rate posts as more offensive when we include age as a categorical variable.

Note that the intercept and effect size are expected to be different because the range of the dependent variable is different for POPQUORN ({1, 2, 3, 4, 5}) and the other two dataset specifications ({0,0.5,1}).

6.2. RQ 2: How does the effect of annotators' demographics on their perception of offensiveness change for posts targeted towards gender?

H2: Annotators will rate content as more offensive if it targets their own in-group compared to out-groups.

Table 4[3] shows the results of the Linear Mixed Effects models applied to three different datasets: Merged, SBIC, and POPQUORN, with offensiveness ratings as the dependent variable. The models examine the influence of annotator demographics (gender, race, age) on the perception of offensiveness in social media posts, comparing posts that target gender versus those that do not.

Merged Dataset

The "Gender" column shows a significant negative coefficient for women and non-binary persons, indicating that women and non-binary annotators tend to rate gender-targeted content as less offensive than men. However, for non-gender-targeted content, women have a non-significant negative coefficient, while gender also shows a non-significant negative coefficient. Among race categories, race Hispanic or Latino, native American, White, and Other have positive and significant coefficients for both gender and non-gender-targeted posts, suggesting these groups rate both types of content as more offensive compared to Asian annotators. The age group 60-64 shows a significant positive coefficient for gender-targeted content, suggesting that older annotators perceive higher offensiveness in these posts.

SBIC Dataset

For gender-targeted content, Woman have a significant negative coefficient, while non-binary persons have a non-significant negative coefficient, indicating that women and non-binary individuals rate gender-targeted posts as less offensive than men. The trend is similar for non-gender-targeted posts, but with less statistical significance. The coefficients for race suggest that Hispanic or Latino and White annotators perceive higher offensiveness in both gender-targeted and non-gender-targeted posts. Older age groups (40-44 and above) tend to rate posts as more offensive, consistent across both gender-targeted and non-gender-targeted categories.

POPQUORN Dataset

In the POPQUORN dataset, Woman and gender non-binary persons show non-significant coefficients for both gender and non-gender-targeted posts, implying that gender does not significantly impact offensiveness ratings in this dataset. The coefficients for Black and White annotators are positive and significant for non-gender-targeted content, indicating these groups rate such content as more offensive compared to Asian annotators. In terms of age, older annotators 60-64 and >65 consistently show positive coefficients, suggesting they rate all types of content as more offensive.

[3] Table 4 can be found in the Appendix D

7. Discussion

Even though our study provides insights into how annotator demographics influence the perception of offensiveness in social media posts, there are several limitations that should be mentioned, which may impact the interpretation of our findings.

One of the key limitations of this study is the demographic bias among annotators, as certain demographic groups within the annotator pool are underrepresented. Specifically, non-binary individuals are significantly underrepresented. In the SBIC dataset, there are only four annotations from non-binary individuals, all of whom rated content with the highest offensiveness scores. This small sample size and high rating could skew our understanding of how non-binary people perceive offensiveness, as it may not be representative of their broader perspective. In contrast, the POPQUORN dataset contains a larger number of non-binary annotations, where their average offensiveness rating is slightly lower than those of men and women. This disparity highlights the potential influence of sample size and dataset composition on our results. As there are significantly more untargeted posts than those explicitly targeting gender, our analysis is also limited by the imbalance in the representation of target groups within the dataset. This imbalance may skew the observed effects when analysing gender-targeted content or other specific groups, potentially leading to less robust conclusions about how different demographics perceive offensiveness in targeted versus untargeted content. The merged dataset used in our analysis is disproportionately composed of posts from the Social Bias Inference Corpus (SBIC), with fewer contributions from the POPQUORN dataset. This imbalance may introduce a dataset-specific bias, where findings are more reflective of the characteristics and annotation guidelines of the SBIC rather than a balanced integration of both datasets.

The process of merging datasets that follow different annotation guidelines presents another challenge. The inconsistencies in how offensiveness was annotated between the SBIC and POPQUORN datasets could lead to variability in the results. These inconsistencies may arise from differences in the interpretation of offensiveness, the context of the posts, or the demographics of the annotators in each dataset, thereby complicating the direct comparison and integration of the data.

8. Validation dictionary method

It is important to validate our dictionary method, because the classification of posts, according to whether they target gender, influences our findings on the second research question. The SBIC has the column *targetCategory* in which annotators also identified to which category the target of each post belonged[4]. The distribution of this column in shown in appendix E. In contrast to the results of our dictionary method, where gender was the most common target category, race is the most common target category in the SBIC annotations.

To evaluate the performance of our dictionary method, we select the categories race, gender and disabled which are present in both our dictionary and the *targetCategory* column. The following steps are described for the posts that target individuals based on their gender but are performed on all three categories in the same way.

[4] Note that each annotator could assign the same post to different categories, so if at least one annotation matched our category of interest we kept that annotation.

First, we separate the annotations that target gender from those that do not. We measure the agreement of the annotators on this division using Krippendorff's alpha. We then label the true target category of the post as gender if half or more of the annotators classify the target of the post as gender, otherwise we assign the label not gender. Finally, we compare this 'true' label with the target detected by the dictionary method (which was also mapped to gender is target or not) and calculate the performance measures.

Table 2: Performance measures

Target category	Agreement	Accuracy	Precision	Recall	F1-score
gender	0.892	0.86	0.368	0.595	0.455
race	0.907	0.894	0.565	0.224	0.321
disabled	0.911	0.98	0.546	0.182	0.273

Table 2 shows the performance measures and Kippendorf's alpha from the mapped TargetCategory column. The inter-annotator agreement is very high for all categories (1 is the highest value of Krippendorf's alpha). Furthermore, the accuracy is relatively high for all categories, but this does not mean that the method detects the target well. The items correctly classified as non-target cause the accuracy to be high, as the distribution between target and non-target is very unbalanced.

Table 3:
LME on subsets based on TargetCategory

	Gender	Not Gender
Intercept	0.953***	0.382***
	(0.009)	(0.004)
gender[T.Woman]	0.003	-0.013***
	(0.004)	(0.002)
gender[T.Non-binary]	0.002	0.125
	(0.138)	(0.197)
race[T.Black]	0.010	0.060***
	(0.014)	(0.007)
race[T.Hispanic or Latino]	-0.001	0.106***
	(0.011)	(0.006)
race[T.Native American]	0.067**	0.076***
	(0.027)	(0.017)
race[T.White]	-0.004	0.065***
	(0.009)	(0.005)
race[T.Other]	-0.017	0.088***
	(0.013)	(0.008)
age[T.25-29]	-0.009	0.045***
	(0.007)	(0.004)
age[T.30-34]	0.016**	-0.005
	(0.007)	(0.004)
age[T.35-39]	0.013*	-0.003
	(0.007)	(0.004)
age[T.40-44]	0.017**	0.079***
	(0.008)	(0.006)
age[T.45-49]	-0.052***	-0.057***
	(0.009)	(0.005)
age[T.50-54]	0.012	0.024**
	(0.014)	(0.010)
age[T.54-59]	-0.003	-0.007
	(0.013)	(0.008)
age[T.60-64]	0.043	0.045
	(0.141)	(0.231)
age[T.>65]	-0.006	0.116***
	(0.032)	(0.033)
Group Var	0.190***	1.958***
	(0.036)	(0.021)
N	7397	105976

Standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01.Reference categories: gender[T.Men], race[T.Asian], age[T.18-24]

Additionally, as can be seen from the other performance measures and the confusion matrices[5], the method does not perform very well. The precision for gender is very low compared to the other categories, but the recall and F1 scores are higher.

To see if the poor performance of our method is driving the results, we split the dataset between “gender” and “not gender” as a target using our “true” label and run the model on both subsets. The results of these regressions are shown in Table 3. In contrast to the results using our dictionary method (shown in Table 4, columns 3 and 4), when gender is the target, the coefficient on women is positive but insignificant. On the other subset, where gender is not the target, the coefficient is negative and significant. These results could be an indication that the low performance of the dictionary method is driving the previous observations. Additionally, the results are consistent with the idea that content targeted at gender, usually women, should be perceived as more offensive by women than content not targeted at them.

[5] The confusion matrices can be found in Appendix F

9. Conclusion

Our study provides insights into how annotator demographics influence the perception of offensiveness in social media posts, with a particular focus on gender-targeted content. While demographics do impact perceptions of offensiveness, the results reveal nuances that diverge from our initial expectations.

Our findings offer partial support for the hypothesis that the perception of offensiveness in social media content is influenced by the demographics of annotators. Notably, annotators older than 60 consistently rated posts as more offensive across datasets. This trend highlights the significance of age in shaping perceptions of offensiveness. However, the demographic influence on offensiveness ratings did not fully align across the datasets. For instance, while the POPQUORN dataset revealed notable racial differences, these did not hold consistently in the SBIC dataset. This discrepancy underscores the complexity of demographic effects and suggests that additional factors may mediate these relationships.

The hypothesis that the impact of gender on the perception of offensiveness is significant for gender-targeted content is partially supported. While the merged dataset shows that women and non-binary annotators tend to rate gender-targeted posts as less offensive compared to men, this pattern was not uniformly observed across all datasets. The SBIC dataset, in particular, demonstrated that women rated gender-targeted posts as less offensive, a result that contrasts with our initial expectations. This finding highlights the need for a more refined approach to understanding how gender-targeted content influences annotator perceptions as the impact of non-binary annotators is limited due to their underrepresentation, making it difficult to draw definitive conclusions about their ratings.

The study's limitations, including the underrepresentation of non-binary annotators and inconsistencies between datasets, affect the robustness of our findings. The dictionary method used for identifying gender-targeted content also proved less effective than anticipated, impacting the analysis of targeted content. Future research should address these limitations by improving target identification methods, balancing demographic representation, and examining additional factors that might influence offensiveness perceptions.

Appendix A:

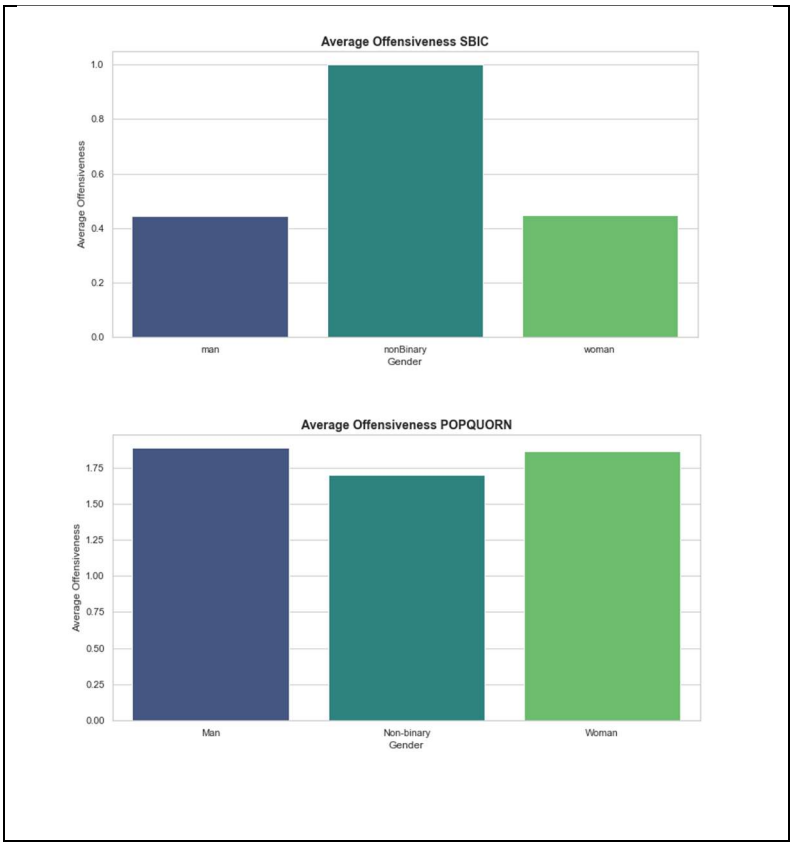

Figure 2: Average Offensiveness Scores for each Dataset

Appendix B:

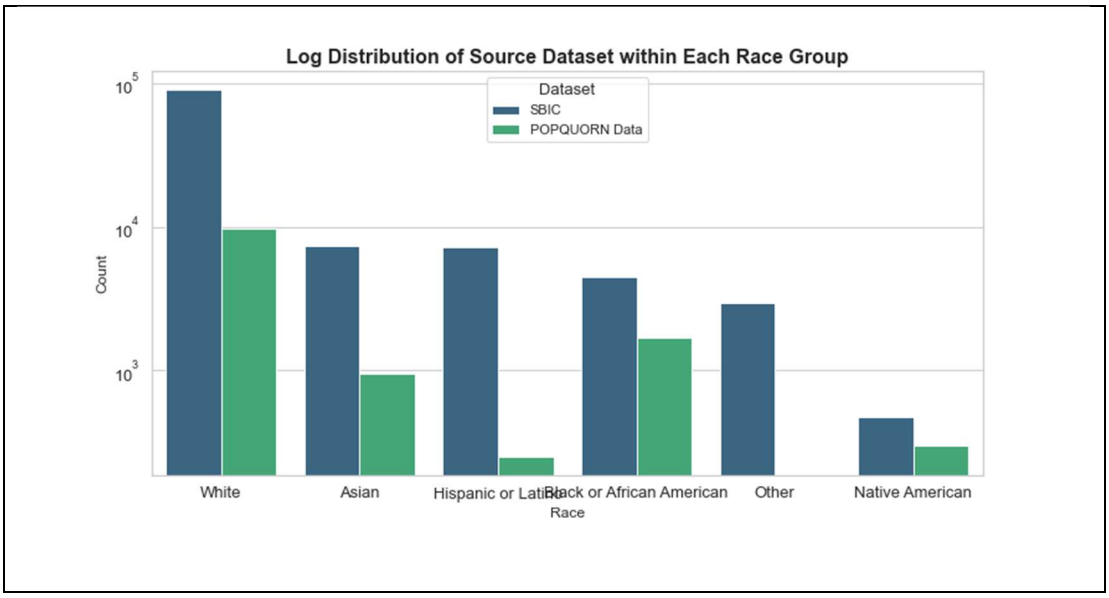

Figure 3: Log Distribution of Source Dataset within Each Race Group

Appendix C:

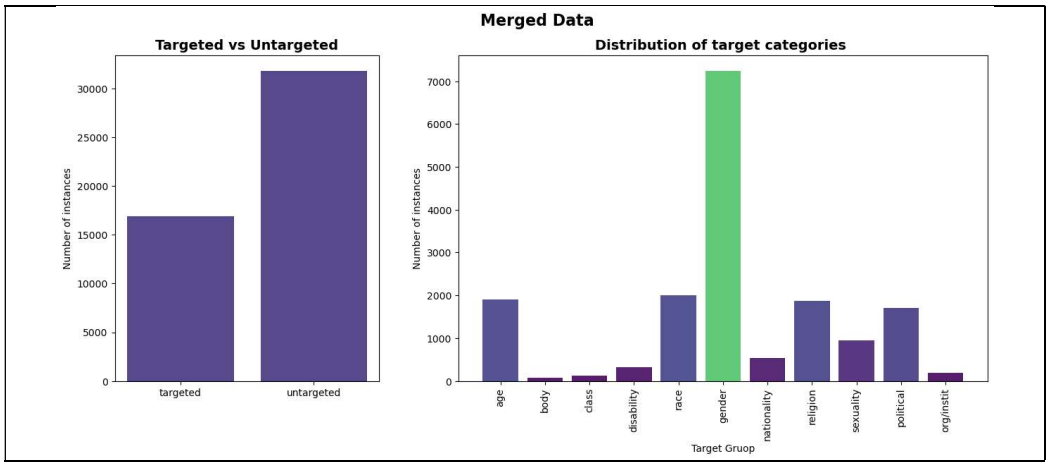

Figure 4: Distribution merged data across targets.

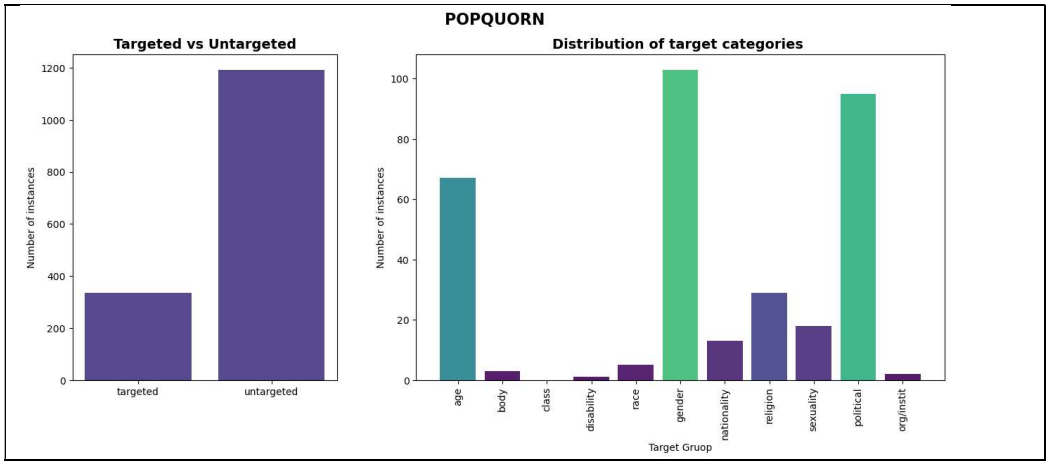

Figure 5: Distribution POPQUORN data across targets.

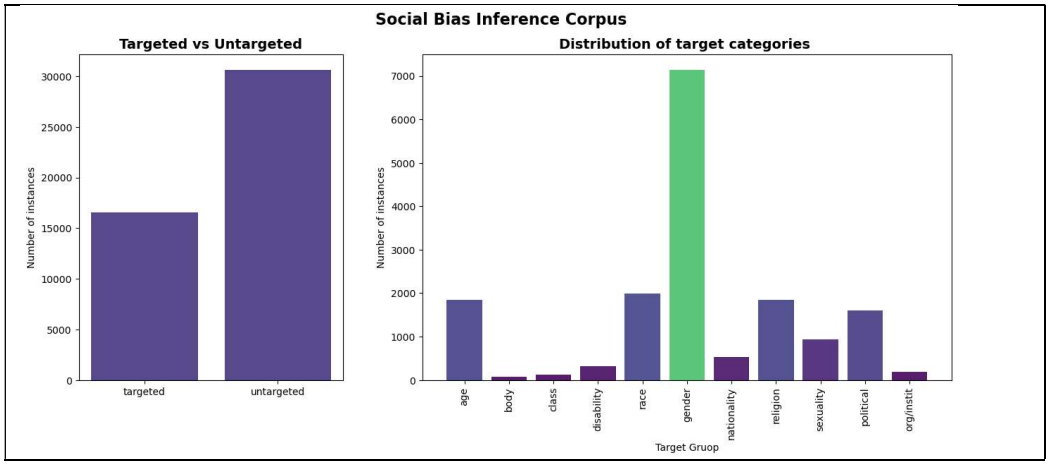

Figure 6: Distribution SBIC data across targets.

Appendix D:

Table 4:
Linear Mixed Effects models applied to subsets of Merged, SBIC, and POPQUORN Datasets

	Merged		SBIC		POPQUORN	
Target	Gender	Not Gender	Gender	Not Gender	Gender	Not Gender
Intercept	0.208***	0.120***	0.587***	0.404***	0.390***	0.204***
	(0.034)	(0.011)	(0.012)	(0.005)	(0.062)	(0.014)
Dataset[T.SBIC]	0.386***	0.294***				
	(0.033)	(0.011)				
gender[T.Woman]	-0.040***	-0.012***	-0.042***	-0.009***	0.018	-0.008
	(0.006)	(0.002)	(0.006)	(0.002)	(0.027)	(0.006)
gender[T.Non-binary]	-0.097	-0.069***		0.097	-0.109	-0.069***
	(0.070)	(0.016)		(0.159)	(0.087)	(0.018)
race[T.Black]	-0.007	0.086***	-0.026	0.067***	-0.033	0.065***
	(0.019)	(0.006)	(0.020)	(0.007)	(0.060)	(0.013)
race[T.Hispanic or Latino]	0.088***	0.092***	0.092***	0.104***	-0.064	-0.088***
	(0.017)	(0.006)	(0.017)	(0.006)	(0.099)	(0.024)
race[T.Native American]	0.124***	-0.001	0.213***	0.071***	-0.177*	-0.116***
	(0.043)	(0.012)	(0.051)	(0.017)	(0.102)	(0.022)
race[T.White]	0.048***	0.055***	0.053***	0.068***	-0.089*	-0.026**
	(0.013)	(0.004)	(0.013)	(0.005)	(0.051)	(0.011)
race[T.Other]	0.076***	0.080***	0.077***	0.089***		
	(0.022)	(0.008)	(0.022)	(0.008)		
age[T.25-29]	0.092***	0.035***	0.099***	0.037***	-0.109**	-0.044***
	(0.011)	(0.004)	(0.011)	(0.004)	(0.055)	(0.012)
age[T.30-34]	0.048***	-0.012***	0.055***	-0.009**	-0.035	-0.033***
	(0.011)	(0.004)	(0.012)	(0.004)	(0.054)	(0.012)
age[T.35-39]	0.067***	-0.009**	0.071***	-0.011***	-0.091*	-0.028**
	(0.011)	(0.004)	(0.011)	(0.004)	(0.054)	(0.012)
age[T.40-44]	0.131***	0.065***	0.131***	0.071***	0.051	0.008
	(0.015)	(0.005)	(0.015)	(0.005)	(0.060)	(0.013)
age[T.45-49]	-0.021	-0.044***	-0.025*	-0.060***	-0.031	0.001
	(0.014)	(0.005)	(0.014)	(0.005)	(0.061)	(0.013)
age[T.50-54]	0.043*	0.017**	0.057**	0.024**	-0.075	-0.006
	(0.025)	(0.007)	(0.028)	(0.010)	(0.063)	(0.013)
age[T.54-59]	0.092***	0.020***	0.090***	-0.018**	0.010	0.036***
	(0.021)	(0.006)	(0.025)	(0.008)	(0.052)	(0.011)
age[T.60-64]	0.169***	0.131***		0.051	0.085	0.112***
	(0.047)	(0.012)		(0.189)	(0.065)	(0.015)
age[T.>65]	0.158***	0.067***	0.239***	0.083***	0.045	0.057***
	(0.032)	(0.009)	(0.081)	(0.032)	(0.053)	(0.012)
Group Var	0.978***	2.036***	1.048***	2.262***	0.191***	0.310***
	(0.030)	(0.022)	(0.032)	(0.025)	(0.046)	(0.017)
N	17178	111884	16277	99694	901	12190

Standard errors in parentheses. *p<0.1, **p<0.05, ***p<0.01
Reference categories: Dataset [POPQUORN], gender[T.Men], race[T.Asian], age[T.18-24]

Appendix E:

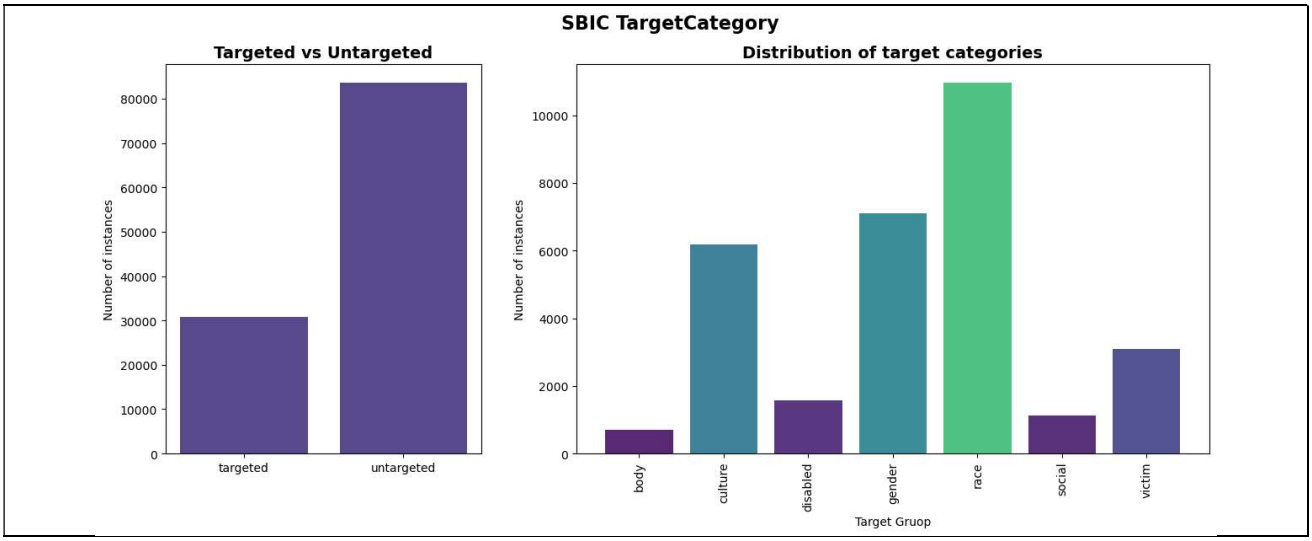

Figure 7: Distribution of *TargetCategory* Column in the SBIC Dataset (Colour depends on the number of instances in each category)

Appendix F:

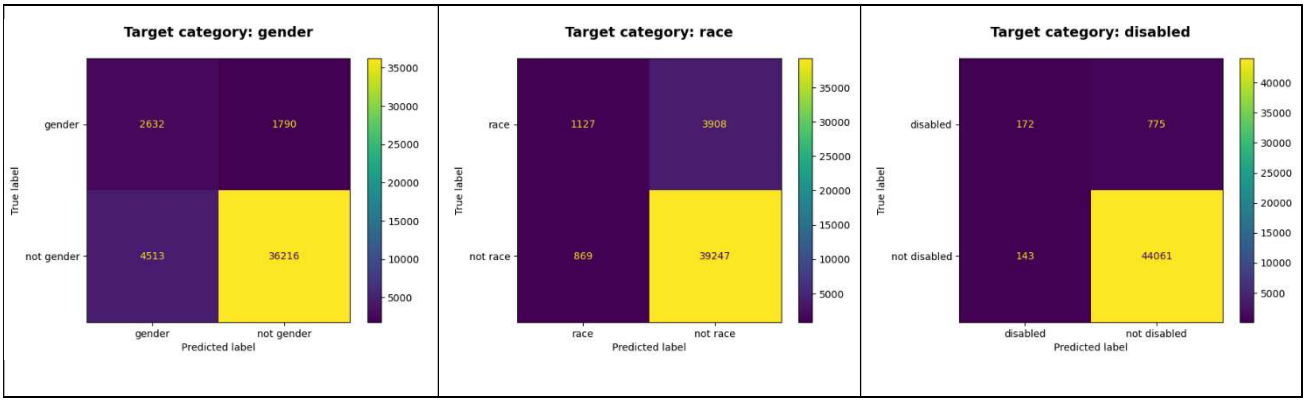

Figure 8: Confusion matrices for validation dictionary method

Autor contributions

The following table shows who contributed for each part of the report and the coding.

Table 5: Author contributions

	Dajana	Ana
Report		
1. Abstract		x
2. Motivation		x
3. Research Background	x	
4. Data	x	
5. Methods		x
6. Results	x	x
6.1. RQ 1		x
6.2. RQ 2	x	
7. Discussion	x	
8. Validation dictionary method		x
9. Conclusion	x	
Code		
PreprocessingData.ipynb	x	
DataExploration.ipynb	x	
DictionaryMethods.ipynb		x
RegressionAnalysis.ipynb		x

x marks the part, which was executed by each of us.

References

Jiang, J., & Nguyen, T. (2007). Linear and generalized linear mixed models and their applications (Vol. 1).

Jiaxin Pei, & David Jurgens. (2023). When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey (2021). Designing toxic content classification for a diversity of perspectives.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Orlikowski, M., Röttger, P., Cimiano, P., & Hovy, D. (2023). The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics.

Perry, T. (2017). Simpledorff - Krippendorff's Alpha On DataFrames. Light Tag. https://www.lighttag.io/blog/krippendorffs-alpha/

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict.

Yu, Z., Sen, I., Assenmacher, D., Samory, M., Fröhling, L., Dahn, C., ... & Wagner, C. (2024). The Unseen Targets of Hate: A Systematic Review of Hateful Communication Datasets

Yiwei Luo, Dallas Card, and Dan Jurafsky (2020) Detecting stance in media on global warming.