Hate Speech Classifiers Learn Normative Social Stereotypes

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, Morteza Dehghani
University of Southern California, USA
{mostafaz,atari,btkenned,mdehghan}@usc.edu

Abstract

Social stereotypes negatively impact individuals' judgments about different groups and may have a critical role in understanding language directed toward marginalized groups. Here, we assess the role of social stereotypes in the automated detection of hate speech in the English language by examining the impact of social stereotypes on annotation behaviors, annotated datasets, and hate speech classifiers. Specifically, we first investigate the impact of novice annotators' stereotypes on their hate-speech-annotation behavior. Then, we examine the effect of normative stereotypes in language on the aggregated annotators' judgments in a large annotated corpus. Finally, we demonstrate how normative stereotypes embedded in language resources are associated with systematic prediction errors in a hate-speech classifier. The results demonstrate that hate-speech classifiers reflect social stereotypes against marginalized groups, which can perpetuate social inequalities when propagated at scale. This framework, combining social-psychological and computational-linguistic methods, provides insights into sources of bias in hate-speech moderation, informing ongoing debates regarding machine learning fairness.

Introduction

Artificial Intelligence (AI) technologies are prone to acquiring cultural, social, and institutional biases from the real-world data on which they are trained (McCradden et al., 2020; Mehrabi et al., 2021; Obermeyer et al., 2019). AI models trained on biased datasets both *reflect* and *amplify* those biases (Crawford, 2017). For example, the dominant practice in modern Natural Language Processing (NLP)—which is to train AI systems on large corpora of human-generated text data—leads to representational biases, such as preferring European American names over African American names (Caliskan et al., 2017), associating words with more negative sentiment with phrases referencing persons with disabilities (Hutchinson et al., 2020), making ethnic stereotypes by associating Hispanics with housekeepers and Asians with professors (Garg et al., 2018), and assigning men to computer programming and women to homemaking (Bolukbasi et al., 2016).

Moreover, NLP models are particularly susceptible to amplifying biases when their task involves evaluating language generated by or describing a social group (Blodgett and O'Connor, 2017). For example, previous research has shown that toxicity detection models associate documents containing features of African American English with higher offensiveness than text without those features (Sap et al., 2019; Davidson et al., 2019). Similarly, Dixon et al. (2018) demonstrate that models trained on social media posts are prone to erroneously classifying ''I am gay'' as hate speech. Therefore, using such models for moderating social-media platforms can yield disproportionate removal of social-media posts generated by or mentioning marginalized groups (Davidson et al., 2019). This unfair assessment negatively impacts marginalized groups' representation in online platforms, which leads to disparate impacts on historically excluded groups (Feldman et al., 2015).

Mitigating biases in hate speech detection, necessary for viable automated content moderation (Davidson et al., 2017; Mozafari et al., 2020), has recently gained momentum (Davidson et al., 2019; Dixon et al., 2018; Sap et al., 2019; Kennedy et al., 2020; Prabhakaran et al., 2019). Most current supervised algorithms for hate speech detection rely on data resources that potentially reflect real-world biases: (1) text representation, which maps textual data to their numeric representations in a semantic space; and (2) human annotations, which represent subjective judgments about the hate speech content of the text, constituting the training dataset. Both (1) and (2) can introduce biases into the final model. First, a classifier may become biased due to how the mapping of language

Transactions of the Association for Computational Linguistics, vol. 11, pp. 300–319, 2023. https://doi.org/10.1162/tacl_a_00550
Action Editor: Alice Oh. Submission batch: 4/2022; Revision batch: 11/2022; Published 3/2023.

to numeric representations is affected by stereotypical co-occurrences in the training data of the language model. For example, a semantic association between phrases referencing persons with disabilities and words with more negative sentiment in the language model can impact a classifier's evaluation of a sentence about disability (Hutchinson et al., 2020). Second, individual-level biases of annotators can impact the classifier in stereotypical directions. For example, a piece of rhetoric about disability can be analyzed and labeled differently depending upon annotators' social biases.

Although previous research has documented stereotypes in text representations (Garg et al., 2018; Bolukbasi et al., 2016; Manzini et al., 2019; Swinger et al., 2019; Charlesworth et al., 2021), the impact of annotators' biases on training data and models remains largely unknown. Filling this gap in our understanding of the effect of human annotation on biased NLP models is the focus of this work. As argued by Blodgett et al. (2020) and Kiritchenko et al. (2021), a comprehensive evaluation of human-like biases in hate speech classification needs to be grounded in social psychological theories of prejudice and stereotypes, in addition to how they are manifested in language. In this paper, we rely on the Stereotype Content Model (SCM; Fiske et al., 2002) which suggests that social perceptions and stereotyping form along two dimensions, namely, *warmth* (e.g., trustworthiness, friendliness) and *competence* (e.g., capability, assertiveness). The SCM's main tenet is that perceived warmth and competence underlie group stereotypes. Hence, different social groups can be positioned in different locations in this two-dimensional space, since much of the variance in stereotypes of groups is accounted for by these basic social psychological dimensions.

In three studies presented in this paper, we study the pipeline for training a hate speech classifier, consisting of collecting annotations, aggregating annotations for creating the training dataset, and training the model. We investigate the effects of social stereotypes on each step, namely, (1) the relationship between social stereotypes and hate speech annotation behaviors, (2) the relationship between social stereotypes and aggregated annotations of trained, expert annotators in curated datasets, and (3) social stereotypes as they manifest in the biased predictions of hate speech classifiers. Our work demonstrates that different stereotypes along warmth and competence differentially affect individual annotators, curated datasets, and trained language classifiers. Therefore, understanding the specific social biases targeting different marginalized groups is essential for mitigating human-like biases of AI models.

1 Study 1: Text Annotation

Here, we investigate the effect of individuals' social stereotypes on their hate speech annotations. Specifically, we aim to determine whether novice annotators' stereotypes (perceived warmth and/or competence) of a mentioned social group lead to higher rate of labeling text as hate speech and higher rates of disagreement with other annotators.

We conduct a study on a nationally stratified sample (in terms of age, ethnicity, gender, and political orientation) of US adults. First, we ask participants to rate eight US-relevant social groups on different stereotypical traits (e.g., friendliness). Then, participants are presented with social media posts mentioning the social groups and are asked to label the content of each post based on whether it attacks the dignity of that group. We expect the perceived warmth and/or competence of the social groups to be associated with participants' annotation behaviors, namely, their rate of labeling text as hate speech and disagreeing with other annotators.

Participants To achieve a diverse set of annotations, we recruited a relatively large (N = 1,228) set of participants in a US sample stratified across participants' gender, age, ethnicity, and political ideology through Qualtrics Panels.[1] After filtering participants based on quality-check items (described below), our final sample included 857 American adults (381 male, 476 female) ranging in age from 18 to 70 (M = 46.7, SD = 16.4) years, about half Democrats (50.4%) and half Republicans (49.6%), with diverse reported race/ethnicity (67.8% White or European American, 17.5% Black or African American, 17.7% Hispanic or Latino/Latinx, 9.6% Asian or Asian American).

[1] https://www.census.gov/quickfacts/fact/table/US/PST045221.

Stimuli To compile a set of stimuli items for this study, we selected posts from the Gab Hate Corpus (GHC; Kennedy et al., 2022), which includes 27,665 social-media posts collected from the corpus of Gab.com (Gaffney, 2018), each annotated for their hate speech content by at least three expert annotators. We collected all posts with high disagreement among the GHC's (original) annotators (based on Equation 1 for quantifying *item disagreement*) which mention at least one social group. We searched for posts mentioning one of the eight most frequently targeted social groups in the GHC: (1) women; (2) immigrants; (3) Muslims; (4) Jews; (5) communists; (6) liberals; (7) African Americans; and (8) homosexual individuals. We selected seven posts per group, resulting in a set of 56 items in total.

Explicit Stereotype Measure We assessed participants' warmth and competence stereotypes of the 8 US social groups in our study based on their perceived traits for a typical member of each group. To this end, we followed social psychological approaches for collecting these self-reported, explicit stereotypes (Cuddy et al., 2008) and asked participants to rate a typical member of each social group (e.g., Muslims) based on their ''friendliness'', ''helpfulness,'' ''peacefulness,'' and ''intelligence.'' Following previous studies of perceived stereotypes (Huesmann et al., 2012; Cuddy et al., 2007), participants were asked to rate these traits from low (e.g., ''unfriendly'') to high (e.g., ''friendly'') using an 8-point semantic differential scale. We considered the average of the first three traits as the indicator of perceived warmth[2] and the fourth item as the perceived competence.

While explicit assessments are generally correlated with implicit measures of attitude, in the case of self-reporting social stereotypes, participants' explicit answers can be less significantly correlated with their implicit biases, potentially due to motivational and cognitive factors (Hofmann et al., 2005). Therefore, it should be noted that this study relies on an explicit assessment of social stereotypes, and the results do not directly explain the effects of *implicit* biases on annotating hate speech.

[2]Cronbach's α's ranged between .90 [women] and .95 [Muslims].

Hate Speech Annotation Task We asked participants to annotate the 56 items based on a short definition of hate speech (Kennedy et al., 2022): ''Language that intends to attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred.''

Participants could proceed with the study only after they acknowledged understanding the provided definition of hate speech. We then tested their understanding of the definition by placing three synthetic ''quality-check'' items among survey items, two of which included clear and explicit hateful language directly matching our definition and one item that was simply informational (see Supplementary Materials). Overall, 371 out of the original 1,228 participants failed to satisfy these conditions and their input was removed from the data.[3]

Disagreement Throughout this paper, we assess annotation disagreement in different levels:

- *Item disagreement,* $d_{(i)}$: Motivated by Fleiss (1971), for each item i, item disagreement $d_{(i)}$ is the number of annotator pairs that disagree on the item's label, divided by the number of all possible annotator pairs.[4]

$$d_{(i)} = \frac{n_1^{(i)} \times n_0^{(i)}}{\binom{n_1^{(i)}+n_0^{(i)}}{2}} \quad (1)$$

  Here, $n_1^{(i)}$ and $n_0^{(i)}$ show the number of hate and non-hate labels assigned to i, respectively.

- *Participant item-level disagreement,* $d_{(p,i)}$: For each participant p and each item i, we define $d_{(p,i)}$ as the ratio of participants with whom p agreed, to the size of the set of participants who annotated the same item (P).

$$d_{(p,i)} = \frac{|\{p' | p' \in P - \{p\}, y_{p,i} = y_{p',i}\}|}{|P|} \quad (2)$$

  Here, $y_{p,i}$ is the label that p assigned to i.

[3]The replication of our analyses with all participants yielded similar results, reported in Supplementary Materials.

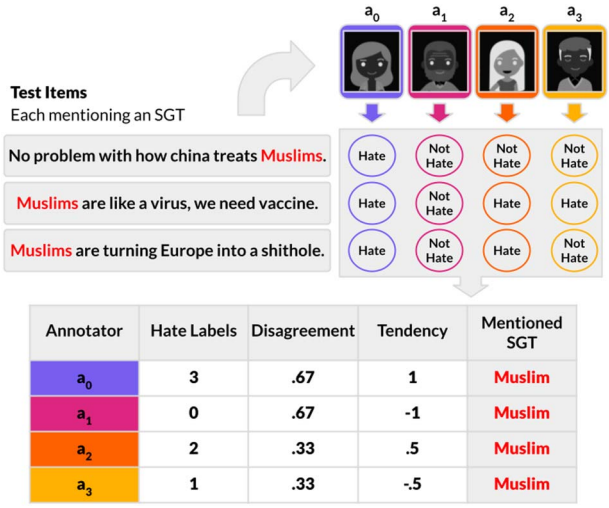
[4]We found this measure more suitable than a simple percentage, as Fleiss captures the total number of annotators as well as the disagreeing pairs.

- *Group-level disagreement,* $d_{(p,S)}$: For a specific set of items S and an annotator p, $d_{(p,S)}$ captures how much p disagrees with others over items in S. We calculate $d_{(p,S)}$ by averaging $d_{(p,i)}$s for all items $i \in S$

$$d_{(p,S)} = \frac{1}{|S|} \sum_{i \in S} d_{(p,i)} \quad (3)$$

Annotators' Tendency To explore participants' annotation behaviors relative to other participants, we rely on the Rasch model (Rasch, 1993). The Rasch model is a psychometric method that models participants' responses—here, annotations—to items by calculating two sets of parameters, namely, the *ability* of each participant and the *difficulty* of each item. Similar approaches, based on Item Response Theory (IRT), have recently been applied in evaluating NLP models (Lalor et al., 2016) and for modeling the relative performance of annotators (Hovy et al., 2013). While, compared to Rasch models, IRT models can include more item-level parameters, our choice of Rasch models is based on their robust estimations for annotators' ability scores. Specifically, Rasch models calculate the ability score solely based on individuals' performances and independent from the sample set. In contrast, in IRT-based approaches, individual annotators' scores depend on the complete set of annotators (Stemler and Naples, 2021). To provide an estimation of these two sets of parameters (annotators' ability and items' difficulty), the Rasch model iteratively fine-tunes parameters' values to ultimately fit the best probability model to participants' responses to items. Here, we apply a Rasch model to each set of items mentioning a specific social group.

It should be noted that Rasch models consider each response as either correct or incorrect and estimate participants' ability and items' difficulty based on the underlying logic that subjects have a higher probability of correctly answering easier items. However, we assume no ''ground truth'' for the labels, therefore ''1''s and ''0''s represent annotators ''hate'' and ''not hate'' answers. Therefore, items' difficulty (which originally represents the probability of ''0'' labels) can be interpreted as *non-hatefulness* (probability of ''non-hate'' labels). Respectively, participants' ability (probability of getting a ''1'' for a difficult item), can be interpreted as their *tendency* towards labeling text as hate (labeling non-hateful items as hateful). Throughout this study we use tendency to refer to the ability parameter.



Annotator	Hate Labels	Disagreement	Tendency	Mentioned SGT
a_0	3	.67	1	Muslim
a_1	0	.67	-1	Muslim
a_2	2	.33	.5	Muslim
a_3	1	.33	-.5	Muslim

Figure 1: The overview of Study 1. Novice annotators are asked to label hate speech content of each post. Then, their annotation behaviors, per social group token, are taken to be the number of posts they labeled as hate speech, their disagreement with other annotators and their tendency to identify hate speech.

Analysis We estimate associations between participants' social stereotypes about each social group with their annotation behaviors evaluated on items mentioning that social group. Namely, the dependent variables are (1) the number of hate labels, (2) the tendency (via the Rasch model) to detect hate speech relative to others, and (3) the ratio of disagreement with other participants—as quantified by *group-level disagreement*. To analyze annotation behaviors concerning each social group, we considered each pair of participant ($N = 857$) and social group ($n_{group} = 8$) as an observation ($n_{total} = 6{,}856$). Each observation includes the social group's perceived warmth and competence based on the participant's answer to the explicit stereotype measure, as well as their annotation behaviors on items that mention that social group. Since each observation is nested in and affected by annotator-level and social-group level variable, we fit cross-classified multi-level models to analyze the association of annotation behaviors with social stereotypes. Figure 1 illustrates our methodology in conducting Study 1. All analyses were performed in

R (3.6.1), and the eRm (1.0.1) package was used for the Rasch model.

Results We first investigated the relation between participants' social stereotypes about each social group and the number of hate speech labels they assigned to items mentioning that group. The result of a cross-classified multi-level Poisson model, with the number of hate speech labels as the dependent variable and participants' perception of warmth and competence as independent variables, shows that a higher number of items are categorized as hate speech when participants perceive that social group as high on competence ($\beta = 0.03, SE = 0.006, p < .001$). In other words, a one point increase in a participant's rating of a social group's competence (on the scale of 1 to 8) is associated with a 3.0% increase in the number of hate labels they assigned to items mentioning that social group. Perceived warmth scores were not significantly associated with the number of hate labels ($\beta = 0.01, p = .128$).

We then compared annotators' relative tendency to assign hate speech labels to items mentioning each social group, calculated by the Rasch models. We conducted a cross-classified multi-level linear model to predict participants' tendency as the dependent variable, and each social group's warmth and competence stereotypes as independent variables. The result shows that participants demonstrate higher tendency (to assign hate speech labels) on items that mention a social group they perceive as highly competent ($\beta = 0.07, SE = 0.013, p < .001$). However, perceived warmth scores were not significantly associated with participants' tendency scores ($\beta = 0.02, SE = 0.014, p = 0.080$).

Finally, we analyzed participants' group-level disagreement for items that mention each social group. We use a logistic regression model to predict disagreement *ratio*, which is a value between 0 and 1. The results of a cross-classified multi-level logistic regression, with group-level disagreement ratio as the dependent variable and warmth and competence stereotypes as independent variables, show that participants disagreed more on items that mention a social group which they perceive as low on competence ($\beta = -0.29, SE = 0.001, p < .001$). In other words, a one point decrease in a participant's rating of a social group's competence (on the scale of 1 to 8) is associated with a 25.2% increase in their odds of disagreement

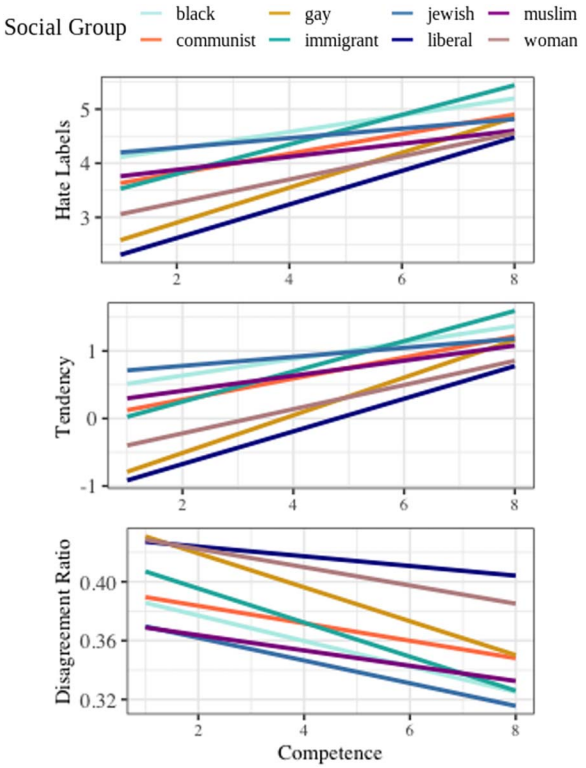


Figure 2: The relationship between the stereotypical competence of social groups and (1) the number of hate labels annotators detected, (2) their tendency to detect hate speech, and (3) their ratio of disagreement with other participants (top to bottom).

on items mentioning that social group. Perceived warmth scores were not significantly associated with the odds of disagreement ($\beta = 0.05, SE = 0.050, p = .322$).

In summary, as represented in Figure 2, the results of Study 1 demonstrate that when novice annotators perceive a mentioned social group as high on competence they (1) assign more hate speech labels, (2) show higher tendency for identifying hate speech, and (3) disagree less with other annotators. These associations collectively denote that when annotators stereotypically perceive a social group as highly competent, they tend to become more sensitive or alert about hate speech directed toward that group. These results support the idea that hate speech annotation is affected by annotators' stereotypes (specifically the perceived competence) of target social groups.

2 Study 2: Ground-Truth Generation

The high levels of inter-annotator disagreements in hate speech annotation (Ross et al., 2017) can

Post	Majority	Disagreement	Warmth	Competence	Mentioned SGT
x_0	0	.67	.16	.12	Muslim
x_1	1	0	.10	.09	Mexican
x_2	1	.67	.16	.18	woman

Figure 3: The overview of Study 2. We investigate a hate speech dataset and evaluate the inter-annotator disagreement and majority label for each document in relation to stereotypes about mentioned social groups.

be attributed to numerous factors, including annotators' varying perception of the hateful language, or ambiguities of the text being annotated (Aroyo et al., 2019). However, aggregating these annotations into single ground-truth labels disregards the nuances of such disagreements (Uma et al., 2021) and even leads to disproportionate representation of individual annotators in annotated datasets (Prabhakaran et al., 2021). Here, we explore the effect of normative social stereotypes, as encoded in language, on the aggregated hate labels provided in a large annotated dataset.

Annotated datasets of hate speech commonly represent the aggregated judgments of annotators rather than individual annotators' annotation behaviors. Therefore, rather than being impacted by individual annotators' self-reported social stereotypes (as in Study 1), we expect aggregated labels to be affected by normative social stereotypes. Here, we rely on semantic representations of social groups in pre-trained language models, known to encode normative social stereotypes and biases of large text corpora (Bender et al., 2021). Figure 3 illustrates the methodology of Study 2.

Data We analyzed the GHC (Kennedy et al., 2022, discussed in Study 1) which includes 27,665 social-media posts labeled for hate speech content by 18 annotators. This dataset includes 91,967 annotations in total, where each post is annotated by at least three coders. Based on our definition of item disagreement in Equation 1, we computed the inter-annotator disagreement and the majority vote for each of the posts and considered them as dependent variables in our analyses.

Quantifying Social Stereotypes To quantify social stereotypes about each social group from our list of social group tokens (Dixon et al., 2018), we calculated semantic similarity of that social group term with lexicons (dictionaries) of competence and warmth (Pietraszkiewicz et al., 2019). The competence and warmth dictionaries consist of 192 and 184 tokens, respectively, and have been shown to measure linguistic markers of competence and warmth reliably in different contexts.

We calculated the similarity of each social group token with the entirety of words in dictionaries of warmth and competence in a latent vector space based on previous approaches (Caliskan et al., 2017; Garg et al., 2018). Specifically, for each social group token, s and each word w in the dictionaries of warmth (D_w) or competence (D_c) we first obtain their numeric representation ($R(s) \in \mathbb{R}^t$ and $R(w) \in \mathbb{R}^t$, respectively) from pre-trained English word embeddings (GloVe; Pennington et al., 2014). The representation function, $R()$, maps each word to a t-dimensional vector, trained based on the word co-occurrences in a corpus of English Wikipedia articles. Then, the warmth and competence scores for each social group token were calculated by averaging the cosine similarity of the numeric representation of the social group token and the numeric representation of the words of the two dictionaries.

Results We examined the effects of the quantified social stereotypes on hate speech annotations captured in the dataset. Specifically, we compared post-level annotation disagreements with the mentioned social group's warmth and competence. For example, based on this method, ''man'' is the most semantically similar social group token to the dictionary of competence ($C_{man} = 0.22$), while ''elder'' is the social group token with the closest semantic representation to the dictionary of warmth ($W_{elder} = 0.19$). Of note, we investigated the effect of these stereotypes on hate speech annotation of social media posts that mention at least one social group token ($N_{posts} = 5535$). Since some posts mention more than one social group token, we considered each mentioned social group token as an observation

($N_{observation} = 7550$), and conducted a multi-level model, with mentioned social group tokens as the level-1 variable and posts as the level-2 variable. We conducted two logistic regression analyses to assess the impact of (1) the warmth and (2) the competence of the mentioned social group as independent variables, and with the inter-annotator disagreement as the dependent variable. The results of the two models demonstrate that both higher warmth ($\beta = -2.62$, $SE = 0.76$, $p < 0.001$) and higher competence ($\beta = -5.27$, $SE = 0.62$, $p < 0.001$) scores were associated with lower disagreement. Similar multi-level logistic regressions with the majority hate label of the posts as the dependent variable and considering either social groups' warmth or competence as independent variables show that competence predicts lower hate ($\beta = -7.77$, $SE = 3.47$, $p = .025$), but there was no significant relationship between perceived warmth and the hate speech content ($\beta = -3.74$, $SE = 4.05$, $p = 0.355$). We like to note that controlling for the frequency of each social groups' mentions in the dataset yields the same results (see Supplementary Materials).

In this study, we demonstrated that social stereotypes (i.e., warmth and competence), as encoded into language resources, are associated with annotator disagreement in an annotated dataset of hate speech. As in Study 1, annotators agreed more on their judgments about social media posts that mention stereotypically more competent groups. Moreover, we observed higher inter-annotator disagreement on social media posts that mentioned stereotypically cold social groups (Figure 4). While Study 1 demonstrated novice annotators' higher tendency for detecting hate speech targeting stereotypically competent groups, we found a lower likelihood of hate labels for posts that mention stereotypically competent social groups in this dataset. The potential reasons for this discrepancy are: (1) while both novice and expert annotators have been exposed to the same definition of hate speech (Kennedy et al., 2018), expert annotators' training focused more on the consequences of hate speech targeting marginalized groups; moreover, the lack of variance in expert annotators' socio-demographic background (mostly young, educated, liberal adults) have led to their increased sensitivity about hate speech directed toward specific stereotypically incompetent groups; and (2) while Study 1 uses a set of items with balanced representation for different

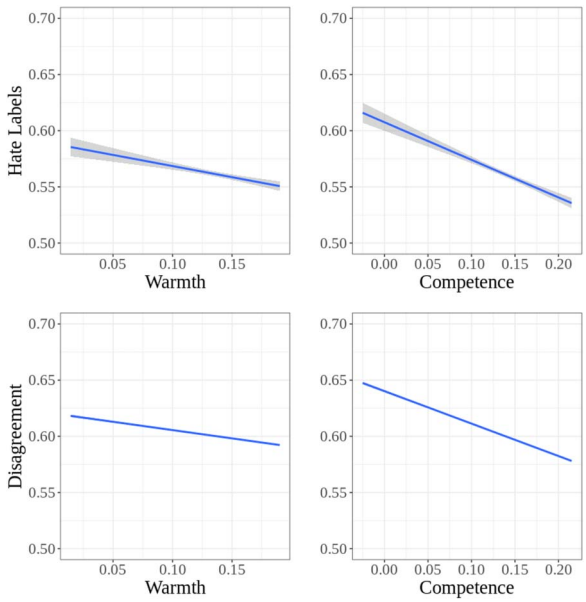


Figure 4: Effects of social groups' stereotype content, on majority hate labels, and annotators' disagreement.

social groups, the dataset used in Study 2 includes disproportionate mentions of social groups. Therefore, the effect might be caused by the higher likelihood of hateful language appearing in GHC's social media posts mentioning stereotypically less competent groups.

3 Study 3: Model Training

NLP models that are trained on human-annotated datasets are prone to patterns of false predictions associated with specific social group tokens (Blodgett and O'Connor, 2017; Davidson et al., 2019). For example, trained hate speech classifiers may have a high probability of assigning a hate speech label to a non-hateful post that mentions the word ''gay.'' Such patterns of false predictions are known as prediction bias (Hardt et al., 2016; Dixon et al., 2018), which impact models' performance on input data associated with specific social groups. Previous research has investigated several sources leading to prediction bias, such as disparate representation of specific social groups in the training data and language models, or the choice of research design and machine learning algorithm (Hovy and Prabhumoye, 2021). However, to our knowledge, no study has evaluated prediction bias with regard to the normative social stereotypes targeting each social group. In Study 3, we investigate whether social stereotypes influence hate speech classifiers' prediction bias toward those groups.

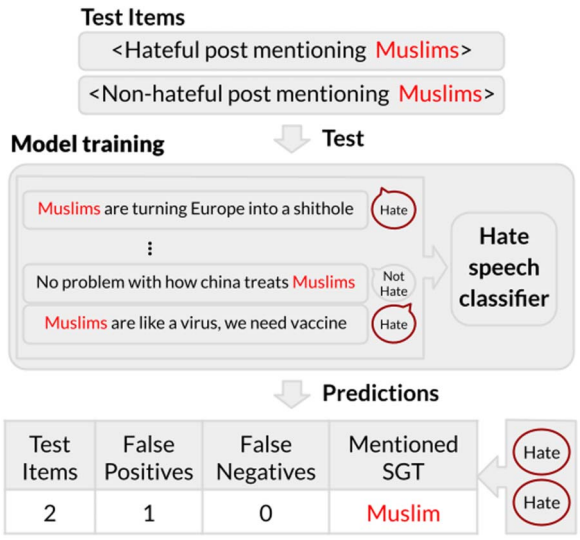


Figure 5: The overview of Study 3. In each iteration, the model is trained on a subset of the dataset. The false predictions of the model are then calculated for each social group token mentioned in test items.

We define prediction bias as erroneous predictions of our text classifier model. We specifically focus on false positives (hate-speech labels assigned to non-hateful instances) and false negatives (non-hate-speech labels assigned to hateful instances) (Blodgett et al., 2020).

In the two previous studies, we demonstrated that variance in annotators' behaviors toward hate speech and imbalanced distribution of ground-truth labels in datasets are both associated with stereotypical perceptions about social groups. Accordingly, we expect hate speech classifiers, trained on the ground-truth labels, to be affected by stereotypes that provoke disagreements among annotators. If that is the case, we expect the classifier to perform less accurately and in a biased way on social-media posts that mention social groups with specific social stereotypes. To detect patterns of false predictions for specific social groups (i.e., prediction bias), we first train several models on different subsets of an annotated corpus of hate speech (GHC; described in Study 1 and 2). We then evaluate the frequency of false predictions provided for each social group and their association with the social groups' stereotypes. Figure 5 illustrates an overview of this study.

Hate Speech Classifiers We implemented three hate speech classifiers; the first two models are based on pre-trained language models, **BERT** (Devlin et al., 2019) and **RoBERTa** (Zhuang et al., 2021). We implemented these two classification models using the `transformers` (v3.1) library of HuggingFace (Wolf et al., 2020) and fine-tuned both models for six epochs with a learning rate of 10^{-7}. The third model applies a Support Vector Machine (**SVM**; Cortes and Vapnik, 1995) with a linear kernel on Term Frequency-Inverse Document Frequency (TF-IDF) vector representations, implemented through the *scikit-learn* (Pedregosa et al., 2011) Python package.

Models were trained on subsets of the GHC and their performance was evaluated on test items mentioning different social groups. To account for possible variations in the resulting models, caused by selecting different subsets of the dataset for training, we performed 100 iterations of model training and evaluating for each classifier. In each iteration, we trained the model on a randomly selected 80% of the dataset ($n_{train} = 22,132$) and recorded the model predictions on the remaining 20% of the samples ($n_{test} = 5,533$). Then, we explored model predictions for all iterations ($n_{prediction} = 100 \times 5,533$), to capture false predictions for instances that mention at least one social group token. By comparing the model prediction with the majority vote for each instance provided in GHC, we detected all ''incorrect'' predictions. For each social group, we specifically capture the number of false-negative (hate speech instances which are labeled as non-hateful) and false-positive (non-hateful instances labeled as hate speech) predictions. For each social group token the false-positive and false-negative ratios are calculated by dividing the number of false predictions by the total number of posts mentioning the social group token.

Quantifying Social Stereotypes In each analysis, we considered either warmth or competence (calculated as in Study 2) of social groups as the independent variable to predict false-positive and false-negative predictions as dependent variables.

Classification Results On average, the classifiers based on BERT, RoBERTa, and SVM achieved F_1 scores of 48.22% ($SD = 3\%$), 47.69% ($SD = 3\%$), and 35.4% ($SD = 1\%$), respectively, on the test sets over the 100 iterations. Since the GHC includes a varying number of posts mentioning each social group token, the

predictions ($n_{prediction} = 553,300$) include a varying number of items for each social group token (M = 2,284.66, Mdn = 797.50, SD = 3,269.20). ''White'' as the most frequent social group token appears in 16,155 of the predictions and ''non-binary'' is the least frequent social group token with only 13 observations. Since social group tokens have varying distributions in the dataset, we considered the ratios of false predictions (rather than frequencies) in all regression models by adding the log-transform of the number of test samples for each social group token as the offset.

Analysis of Results The average false-positive ratio of social group tokens in the BERT-classifier was 0.58 ($SD = 0.24$), with a maximum of 1.00 false-positive ratio for several social groups, including ''bisexual'', and the minimum of 0.03 false-positive ratio for ''Buddhist.'' In other words, BERT-classifiers always predicted incorrect hate speech labels for non-hateful social-media posts mentioning ''bisexuals'' while rarely making those mistakes for posts mentioning ''Buddhists''. The average false-negative ratio of social group tokens in the BERT-classifier was 0.12 ($SD = 0.11$), with a maximum of 0.49 false-negative ratio associated with ''homosexual'' and the minimum of 0.0 false-negative ratio for several social groups including ''Latino.'' In other words, BERT-classifiers predicted incorrect non-hateful labels for social-media post mentioning ''homosexuals'' while hardly making those mistakes for posts mentioning ''Latino''. These statistics are consistent with observations of previous findings (Davidson et al., 2017; Kwok and Wang, 2013; Dixon et al., 2018; Park et al., 2018), which identify false-positive errors as the more critical issue with hate speech classifiers.

For each classifier, we assess the number of false-positive and false-negative hate speech predictions for social-media posts that mention each social group. For analyzing each classifier, two Poisson models were created, considering false-positive predictions as the dependent variable and social groups' (1) warmth or (2) competence, calculated from a pre-trained language model (see Study 2) as the independent variable. The same settings were considered in two other Poisson models to assess false-negative predictions as the dependent variable, and either warmth or competence as the independent variable.

	False Positive		False Negative	
	W	C	W	C
BERT	−0.09**	−0.23**	−0.04**	−0.10**
RoBERTa	−0.04**	−0.15**	0.02	0.05*
SVM	−0.05**	−0.09**	−0.01	−0.09**

Table 1: Associations of erroneous predictions (false positives and false negatives) and social groups' warmth (W) and competence (C) stereotypes in predictions of three classifiers. ** and * represent p-values less than .001 and .05, respectively.

Table 1 reports the association between social groups' warmth and competence stereotypes with the false hate speech labels predicted by the models. The results indicate that the number of false-positive predictions is negatively associated with the social groups' language-embedded warmth and competence scores in all three models. In other words, texts that mentions social groups stereotyped as cold and incompetent are more likely to be misclassified as containing hate speech; for instance, in the BERT-classifier a one point increase in the social groups warmth and competence is, respectively, associated with 8.4% and 20.3% decrease in model's false-positive error ratios. The number of false-negative predictions is also significantly associated with the social groups' competence scores; however, this association had varying directions among the three models. BERT and SVM classifiers are more likely to misclassify instances as not containing hate speech when texts mention stereotypically incompetent social groups; such that one point increase in competence is associated with 9.8% decrease in BERT model's false-negative error ratio. Whereas false-negative predictions of the RoBERTa model is more likely for text mentioning stereotypically competent social groups. The discrepancy in the association of warmth and competence stereotypes and false-negative errors calls for further investigation. Figure 6 depicts the associations of the two stereotype dimensions with the proportions of false-positive and false-negative predictions of the BERT classifier for social groups.

In summary, this study demonstrates that erroneous predictions of hate speech classifiers are associated with the normative stereotypes regarding the social groups mentioned in text.

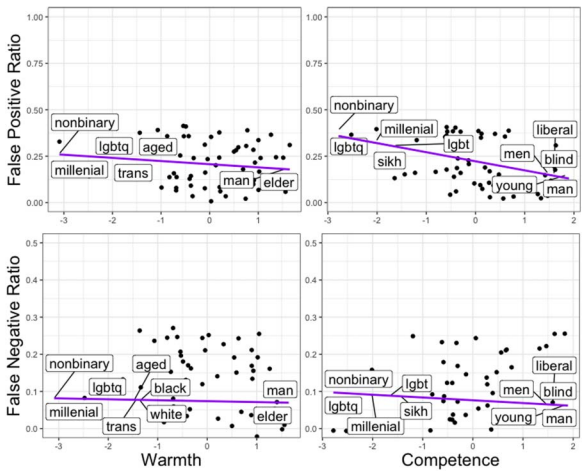


Figure 6: Social groups' higher stereotypical competence and warmth is associated with lower false positive and negative predictions in hate speech detection.

Particularly, the results indicate that documents mentioning stereotypically colder and less competent social groups, which lead to higher disagreement among expert annotators based on Study 2, drive higher error rates in hate speech classifiers. This pattern of high false predictions (both false-positives and false-negatives) for social groups stereotyped as cold and incompetent implies that prediction bias in hate speech classifiers is associated with social stereotypes, and resembles normative social biases that we documented in the previous studies.

4 Discussion

Here, we integrate theory-driven and data-driven approaches (Wagner et al., 2021) to investigate human annotators' and normative social stereotypes as a source of bias in hate speech datasets and classifiers. In three studies, we combine social psychological frameworks and computational methods to make theory-driven predictions about hate-speech-annotation behavior and empirically test the sources of bias in hate speech classifiers. Overall, we find that hate speech annotation behaviors, often assumed to be objective, are impacted by social stereotypes, and that this in turn adversely influences automated content moderation.

In Study 1, we investigated the association between participants' self-reported social stereotypes against 8 different social groups, and their annotation behavior on a small subset of social-media posts about those social groups. Our findings indicate that for novice annotators judging social groups as competent is associated with a higher tendency toward detecting hate and lower disagreement with other annotators. We reasoned that novice annotators prioritize protecting the groups they perceive as warm and competent. These results can be interpreted based on the Behaviors from Intergroup Affect and Stereotypes framework (BIAS; Cuddy et al., 2007): groups judged as competent elicit passive facilitation (i.e., obligatory association), whereas those judged as lacking competence elicit passive harm (i.e., ignoring). Here, novice annotators might tend to ''ignore'' social groups judged to be incompetent and not assign ''hate speech'' labels to inflammatory posts attacking these social groups.

However, Study 1's results may not uncover the pattern of annotation biases in hate speech datasets as data curation efforts rely on annotator pools with imbalanced representation of different socio-demographic groups (Posch et al., 2018) and data selection varies among different datasets. In Study 2, we examined the role of social stereotypes in the aggregation process, where expert annotators' disagreements are discarded to create a large dataset containing the ground-truth hate-speech labels. We demonstrated that, similar to Study 1, texts that included groups stereotyped to be warm and competent were highly agreed upon. However, unlike Study 1, posts mentioning groups stereotyped as incompetent are more frequently marked as hate speech by the aggregated labels. In other words, novice annotators tend to focus on protecting groups they perceive as competent; however, the majority vote of expert annotators tend to focus on common targets of hate in the corpus. We noted two potential reasons for this disparity (1) Novice and expert annotators vary in their annotation behaviors; in many cases, hate speech datasets are labeled by expert annotators who are thoroughly trained for this specific task (Patton et al., 2019), and have specific experiences that affect their perception of online hate (Talat, 2016). GHC annotators were undergraduate psychologist research assistants trained by first reading a typology and coding manual for studying hate-based rhetoric and then passing a curated test of about thirty messages designed for assessing their understanding of the annotation task (Kennedy et al., 2022). Therefore, their relatively higher familiarity with and experience in

annotating hate speech, compared to annotators in Study 1, led to different annotation behaviors. Moreover, dataset annotators are not usually representative of the exact population that interacts with social media content. As pointed out by Díaz et al. (2022), understanding the socio-cultural factors of an annotator pool can shed light on the disparity of our results. In our case, identities and lived experiences can significantly vary between participants in Study 1 and GHC's annotators in Study 2, which impacts how annotation questions are interpreted and responded to. (2) Social groups with specific stereotypes have imbalanced presence in hate speech datasets; while in Study 1, we collect a balanced set of items with equal representation for each of the 8 social groups, social media posts disproportionately include mentions of different social groups, and the frequency of each social group being targeted depends on multiple social and contextual factors.

To empirically demonstrate the effect of social stereotypes on supervised hate speech classifiers, in Study 3, we evaluated the performance and biased predictions of such models when trained on an annotated dataset. We used the ratio of incorrect predictions to operationalize the classifiers' unintended bias in assessing hate speech toward specific groups (Hardt et al., 2016). Study 3's findings suggested that social stereotypes of a mentioned group, as captured in large language models, are significantly associated with biased classification of hate speech such that more false-positive predictions are generated for documents that mention groups that are stereotyped to be cold and incompetent. However, we did not find consistent trends in associations between social groups' warmth and competence stereotypes and false-negative predictions among different models. These results demonstrate that false-positive predictions are more frequent for the same social groups that evoked more disagreements between annotators in Study 2. Similar to Davani et al. (2022), these findings challenge supervised learning approaches that only consider the majority vote for training a hate speech classifier and dispose of the annotation biases reflected in inter-annotator disagreements.

It should be noted that while Study 1 assesses social stereotypes as reported by novice annotators, Studies 2 and 3 rely on a semantic representation of such stereotypes. Since previous work on language representation have shown that semantic representations encode socially embedded biases, in Studies 2 and 3 we referred to the construct under study as normative social stereotypes. Our comparison of results demonstrated that novice annotators' self-reported social stereotypes impact their annotation behaviors, and the annotated datasets and hate speech classifiers are prone to being affected by normative stereotypes.

Our work is limited to the English language, a single dataset of hate speech, and participants from the US. Given that the increase in hate speech is not limited to the US, it is important to extend our findings in terms of research participants and language resources. Moreover, we applied SCM to quantify social stereotypes, but other novel theoretical frameworks such as the Agent-Beliefs-Communion model (Koch et al., 2016) can be applied in the future to uncover other sources of bias.

5 Related Work

Measuring Annotator Bias Annotators are biased in their interpretations of subjective language understanding tasks (Aroyo et al., 2019; Talat et al., 2021). Annotators' sensitivity to toxic language can vary based on their expertise (Talat, 2016), lived experiences (Patton et al., 2019), and demographics (e.g., gender, race, and political orientation) (Cowan et al., 2002; Norton and Sommers, 2011; Carter and Murphy, 2015; Prabhakaran et al., 2021; Jiang et al., 2021). Sap et al. (2022) discovered associations between annotators' racist beliefs and their perceptions of toxicity in anti-Black messages and text written in African American English. Compared to previous efforts, our research takes a more general approach to modeling annotators' biases, which is not limited to specific targets of hate.

Recent research efforts argue that annotators' disagreements should not be treated solely as noise in data (Pavlick and Kwiatkowski, 2019) and call for alternative approaches for considering annotators as independent sources for informing the modeling process in subjective tasks (Prabhakaran et al., 2021). Such efforts tend to improve data collection (Vidgen et al., 2021; Rottger et al., 2022) and the modeling process in various tasks, such as detecting sarcasm (Rajadesingan et al., 2015), humor (Gultchin et al., 2019), sentiment (Gong et al., 2017), and hate speech (Kocoń et al., 2021). For instance, Davani et al. (2022)

introduced a method for modeling individual annotators' behaviors rather than their majority vote. In another work, Akhtar et al. (2021) clustered annotators into groups with high internal agreement (similarly explored by Wich et al., 2020) and redefined the task as modeling the aggregated label of each group. Our findings especially help such efforts by providing a framework for incorporating annotators' biases into hate speech classifiers.

Measuring Hate Speech Detection Bias When propagated into the modeling process, biases in the annotated hate speech datasets cause group-based biases in predictions (Sap et al., 2019) and lack of robustness in results (Geva et al., 2019; Arhin et al., 2021). Specifically, previous research has shed light on *unintended* biases (Dixon et al., 2018), which are generally defined as systemic differences in performance for different demographic groups, potentially compounding existing challenges to fairness in society at large (Borkan et al., 2019). While a significant body of work has been dedicated to mitigating unintended biases in hate speech (and abusive language) classification (Vaidya et al., 2020; Ahmed et al., 2022; Garg et al., 2019; Nozza et al., 2019; Badjatiya et al., 2019; Park et al., 2018; Mozafari et al., 2020; Xia et al., 2020; Kennedy et al., 2020; Mostafazadeh Davani et al., 2021; Chuang et al., 2021), the choice of the exact bias metrics is not consistent within all these studies. As demonstrated by Czarnowska et al. (2021), various bias metrics can be considered as different parametrizations of a generalized metric. In hate speech detection in particular, disproportionate false predictions, especially false positive predictions, for marginalized social groups have often been considered as an indicator of unintended bias in the model. This is due to the fact that hate speech, by definition, involves a social group as the target of hate, and the disproportionate mentions of specific social groups in hateful social media content have led to imbalance datasets and biased models.

Measuring Social Stereotypes The Stereotype Content Model (SCM; Fiske et al., 2002) suggests that to determine whether other people are threats or allies, individuals make prompt assessments about their warmth (good vs. ill intentions) and competence (ability vs. inability to act on intentions). Koch et al. (2016) proposed to fill in an empirical gap in SCM by introducing the *ABC model* of stereotype content. Based on this model, people organize social groups primarily based on their (A) agency (competence in SCM), and (B) conservative-progressive beliefs. They did not find (C) communion (warmth in SCM) as a dimension by its own, but rather as an emergent quality in the other two dimensions. Zou and Cheryan (2017) proposed that racial and ethnic minority groups are disadvantaged along two distinct dimensions of perceived inferiority and perceived cultural foreignness, which can explain the patterns of racial prejudice.

Among different social psychological frameworks for social stereotypes, we relied on SCM in this research, as it has been shown helpful for predicting emotional and behavioral reactions to outgroups. For instance, the Behaviors from Intergroup Affect and Stereotypes framework (BIAS; Cuddy et al., 2007), an extension of the SCM, explains how stereotypes shape behavioral tendencies toward groups and individuals (Cuddy et al., 2008). Based on this theoretical framework, perceived warmth predicts *active* behaviors, while perceived competence predicts *passive* behaviors toward outgroups. Individuals tend to express admiration for social groups stereotyped as warm and competent and facilitate them both actively (by protecting them) and passively (by cooperating with them). In contrast, people tend to express hate toward social groups stereotyped as cold and incompetent and harm them actively (by attacking) and passively (by excluding).

6 Conclusion

Our findings suggest that hate speech classifiers trained on human annotations will acquire particular prediction biases toward historically marginalized groups. Our results have two specific and direct implications: First, one way to decrease unintended bias in classifiers is to study and diversify annotation teams, so that annotators come from different ideological and social backgrounds; by doing so, coders may not agree on a label to (over)protect an ingroup or ''ally'' groups (leading to false positives), or to ignore actually hateful content targeting ''blameworthy'' groups (leading to false negatives). Second, supervised learning approaches may benefit from modeling annotation biases, which are reflected in inter-annotator disagreements, rather than the current practice,

which is to treat them as unexplained noise in human judgment, to be disposed of through annotation aggregation. Our work is an example of how well-established social scientific theories can be applied to explain the novel interactions between algorithms and people. Large amounts of data that are being constantly recorded in ever-changing socio-technical environments call for integrating social scientific theories for assessing the novel technologies and their associated social impacts.

A Supplementary Materials

Study 1 – Test Items The three test items are: (1) I want to see Sandirians raped to death collectively. (2) Now if we could only get a few jew heads on pikes that would be great. (3) Over 99% of Syrian refugees admitted to UK in 2017 were Muslims. We expect annotators to label (1) and (2) as **hate speech** and label (3) as **not hate speech**.

Study 1 – Analysis of All Annotators We replicate the results of Study 1, on the whole set of participants (N = 1,228). The result shows that a higher number of items are categorized as hate speech when participants perceive that social group as high on competence ($\beta = 0.02$, $SE = 0.005$, $p < .001$). However, warmth scores were not significantly associated with the number of hate-speech labels ($\beta = 0.01$, $SE = 0.006$, $p = .286$). Moreover, participants demonstrate higher tendency (to assign hate speech labels) on items that mention a social group they perceive as highly competent ($\beta = 0.04$, $SE = 0.010$, $p < .001$). Warmth scores were only marginally associated with participants' *tendency* scores ($\beta = 0.02, SE = 0.010, p = 0.098$). Lastly, participants disagreed more on items that mention a social group perceived as incompetent ($\beta = -0.17$, $SE = 0.034$, $p <.001$). Contrary to the original results, warmth scores were also significantly associated with the odds of disagreement ($\beta = 0.07, SE = 0.036, p = .044$).

Study 1 and 2 – Stereotypes Table 2 reports the calculated stereotype scores for each social group.

Study 2 assesses over 63 social groups; the calculated warmth score varies from 0.01 to 0.19 (mean = 0.14, sd = 0.03), and competence varies from −0.03 to 0.22 (mean = 0.14, sd = 0.04).

	Study 1		Study 2	
Group	C	W	C	W
Immigrant	6.8 (1.6)	7.2 (1.4)	5.0	5.0
Muslim	7.0 (1.7)	6.6 (1.8)	4.9	5.1
Communist	5.8 (2.0)	5.1 (2.0)	5.0	5.0
Liberal	6.7 (2.0)	6.6 (1.9)	5.2	5.1
Black	7.0 (1.7)	6.9 (1.6)	4.8	4.7
Gay	7.3 (1.5)	7.5 (1.4)	4.9	5.1
Jewish	7.7 (1.3)	7.3 (1.4)	4.9	5.0
Woman	7.6 (1.3)	7.5 (1.2)	5.2	5.1

Table 2: Perceived warmth (W) and competence (C) scores varying from 1 (most negative trait) to 8 (most positive trait). Study 1 columns represent the average and standard deviation of participants' responses. Study 2's values are scaled from [−1, 1] to [1, 8]. The correlation of perceived competence and warmth score within the two studies are −0.07 and 0.09, respectively.

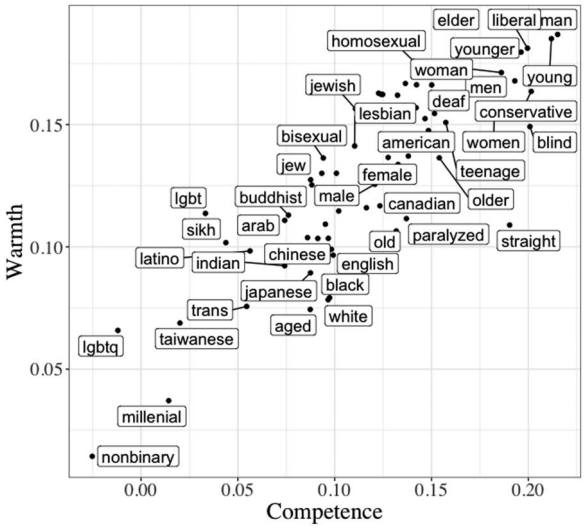


Figure 7: The distribution of social groups on the warmth-competence space based the calculated scores used in Study 2.

Figure 7 plots the social groups on the warmth and competence dimensions calculated in Study 2.

Study 2 – Frequency as a Control Variable After adding social groups' frequency as a control variable, both higher warmth ($\beta = -2.28$, $SE = 0.76$, $p < 0.01$) and competence ($\beta = -5.32$, $SE = 0.62$, $p < 0.001$) scores were associated with lower disagreement. Competence predicts lower hate ($\beta = -7.96$, $SE = 3.71$, $p = .032$), but there was no significant relationship between perceived warmth and the hate speech content ($\beta = -2.95, SE = 3.89, p = .448$).

Acknowledgments

We would like to thank Nils Karl Reimer, Vinodkumar Prabhakaran, Stephen Read, the anonymous reviewers, and the action editor for their suggestions and feedback.

References

Zo Ahmed, Bertie Vidgen, and Scott A. Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8. https://doi.org/10.1140/epjds/s13688-022-00319-9

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-truth, whose truth?–examining the challenges with annotating toxic text datasets. *arXiv preprint arXiv:2112.03529*.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1100–1105. https://doi.org/10.1145/3308560.3317083

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of ‘‘bias’’ in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Su Lin Blodgett and Brendan O’Connor. 2017. Racial disparity in natural language processing: A case study of social media African American English. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, KDD*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 491–500.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

Evelyn R. Carter and Mary C. Murphy. 2015. Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass*, 9(6):269–280. https://doi.org/10.1111/spc3.12181

Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32:218–240. https://doi.org/10.1177/0956797620963619, PubMed: 33400629

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen,

and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 114–120, Online. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297. https://doi.org/10.1007/BF00994018

Gloria Cowan, Miriam Resendez, Elizabeth Marshall, and Ryan Quist. 2002. Hate speech and constitutional protection: Priming values of equality and freedom. *Journal of Social Issues*, 58(2):247–263. https://doi.org/10.1111/1540-4560.00259

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2007. The bias map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 92(4): 631–648. https://doi.org/10.1037/0022-3514.92.4.631, PubMed: 17469949

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in Experimental Social Psychology*, 40:61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267. https://doi.org/10.1162/tacl_a_00425

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110. https://doi.org/10.1162/tacl_a_00449

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3504

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351. https://doi.org/10.1145/3531146.3534647

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. https://doi.org/10.1145/3278721.3278729

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Susan T. Fiske, Amy J. C. Cuddy, P. Glick, and J. Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878. https://doi.org/10.1037/0022-3514.82.6.878

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Gavin Gaffney. 2018. Pushshift gab corpus. https://files.pushshift.io/gab/. Accessed: 2019-5-23.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. https://doi.org/10.1073/pnas.1720347115, PubMed: 29615513

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. https://doi.org/10.1145/3306618.3317950

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1107

Lin Gong, Benjamin Haines, and Hongning Wang. 2017. Clustered model adaption for personalized sentiment analysis. In *Proceedings of the 26th International Conference on World Wide Web*, pages 937–946.

Limor Gultchin, Genevieve Patterson, Nancy Baym, Nathaniel Swinger, and Adam Kalai. 2019. Humor in word embeddings: Cockamamie gobbledegook for nincompoops. In *International Conference on Machine Learning*, pages 2474–2483. PMLR.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Wilhelm Hofmann, Bertram Gawronski, Tobias Gschwendner, Huy Le, and Manfred Schmitt. 2005. A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31(10):1369–1385. https://doi.org/10.1177/0146167205275613, PubMed: 16143669

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432. https://doi.org/10.1111/lnc3.12432, PubMed: 35864931

L. Rowell Huesmann, Eric F. Dubow, Paul Boxer, Violet Souweidane, and Jeremy Ginges. 2012. Foreign wars and domestic prejudice: How media exposure to the Israeli-Palestinian conflict predicts ethnic stereotyping by Jewish and Arab American adolescents. *Journal of Research on Adolescence*, 22(3):556–570. https://doi.org/10.1111/j.1532-7795.2012.00785.x, PubMed: 23243381

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS One*, 16(8):e0256762. https://doi.org/10.1371/journal.pone.0256762, PubMed: 34449815

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

https://doi.org/10.1007/s10579-021-09569-x

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.483

Brendan Kennedy, Drew Kogon, Kris Coombs, Joseph Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh Davani, Mohammad Atari, and Morteza Dehghani. 2018. A typology and coding manual for the study of hate-based rhetoric. *PsyArXiv. July*, 18.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478. https://doi.org/10.1613/jair.1.12590

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The abc of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675. https://doi.org/10.1037/pspa0000046, PubMed: 27176773

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173. IEEE.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27. https://doi.org/10.1609/aaai.v27i1.8539

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648. NIH Public Access.

Thomas Manzini, Lim Yao Chong, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1062

Melissa D. McCradden, Shalmali Joshi, James A. Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. 2020. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, 27(12):2024–2027. https://doi.org/10.1093/jamia/ocaa085, PubMed: 32585698

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35. https://doi.org/10.1145/3457607

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS ONE*, 15(8):e0237861. https://doi.org/10.1371/journal.pone.0237861, PubMed: 32853205

Michael I. Norton and Samuel R. Sommers. 2011. Whites see racism as a zero-sum game that they are now losing. *Perspectives on Psychological Science*, 6(3):215–218. https://doi.org/10.1177/1745691611406922, PubMed: 26168512

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny

detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155. https://doi.org/10.1145/3350546.3352512

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453. https://doi.org/10.1126/science.aax2342, PubMed: 31649194

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2019.260

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Agnieszka Pietraszkiewicz, Magdalena Formanowicz, Marie Gustafsson Sendén, Ryan L. Boyd, Sverker Sikström, and Sabine Sczesny. 2019. The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, 49(5):871–887. https://doi.org/10.1002/ejsp.2561

Lisa Posch, Arnim Bleier, Fabian Flöck, and Markus Strohmaier. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. In *Eleventh International AAAI Conference on Web and Social Media*.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1578

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106. https://doi.org/10.1145/2684822.2685316

Georg Rasch. 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. ERIC.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer Mediated Communication*.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting

data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.13

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.

Steven E. Stemler and Adam Naples. 2021. Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation*, 26:11.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311. https://doi.org/10.1145/3306618.3314270

Zeerak Talat. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. Anonymous preprint under review.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470. https://doi.org/10.1613/jair.1.12752

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.132

Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. 2021. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204. https://doi.org/10.1038/s41586-021-03666-1, PubMed: 34194046

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the fourth workshop on online abuse and harms*, pages 191–199.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*,

pages 7–14, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Linda X. Zou and Sapna Cheryan. 2017. Two axes of subordination: A new model of racial position. *Journal of Personality and Social Psychology*, 112(5):696–717. https://doi.org/10.1037/pspa0000080, PubMed: 28240941