Targeted Identity Group Prediction in Hate Speech Corpora

Pratik S. Sachdeva
D-Lab
University of California, Berkeley
pratik.sachdeva@berkeley.edu

Renata Barreto
School of Law
University of California, Berkeley
rbarreto@berkeley.edu

Claudia von Vacano
D-Lab
University of California, Berkeley
cvacano@berkeley.edu

Chris J. Kennedy
Center for Precision Psychiatry
Harvard Medical School
chris_kennedy@hms.harvard.edu

Abstract

The past decade has seen an abundance of work seeking to detect, characterize, and measure online hate speech. A related, but less studied problem, is the specification of identity groups targeted by that hate speech. Predictive accuracy on this task can supplement additional analyses beyond hate speech detection, motivating its study. Using the *Measuring Hate Speech* corpus, which provided annotations for targeted identity groups on roughly 50,000 social media comments, we create neural network models to perform multi-label binary prediction of identity groups targeted by a social media comment. Specifically, we study 8 broad identity groups and 12 identity sub-groups within race and gender identity. We find that these networks exhibited good predictive performance, achieving ROC AUCs of greater than 0.9 and PR AUCs of greater than 0.7 on several identity groups. At the same time, we find performance suffered on identity groups less represented in the dataset. We validate model performance on the HateCheck and Gab Hate Corpora, finding that predictive performance generalizes in most settings. We additionally examine the performance of the model on comments targeting multiple identity groups. Lastly, we discuss issues with a standardized conceptualization of a "target" in hate speech corpora, and its relation to intersectionality. Our results demonstrate the feasibility of simultaneously detecting a broad range of targeted groups in social media comments, and offer suggestions for future work on modeling and dataset annotation for this task.

1 Introduction

The proliferation of hate speech on online platforms continues to be a significant human rights issue, associated with a host a negative consequences (Tsesis, 2002; Wilson, 2017). Hate speech distinguishes itself from other types of toxic or offensive content in that it specifically targets an individual or group on the basis of their membership in an identity group, such as race, religion, gender, sexual orientation, etc. (Sellars, 2016). Thus, developing methods that can identify and characterize hate speech, and its targets, is of paramount importance.

Given the scale of online hate speech, much effort has been made toward the development of automated approaches to classify or measure it given raw text (Fortuna and Nunes, 2018; Tontodimamma et al., 2021). While initial efforts used binary labels, subsequent work has introduced additional labels that more finely characterize or measure hate speech (Kennedy et al., 2020; Davidson et al., 2017; Kennedy et al., 2022). These include studies that implicitly specify the targeted identity group, such as labeling speech as racism or sexism (Waseem and Hovy, 2016).

Predicting the identity group targeted by social media content is useful beyond hate speech detection. Such algorithms could identify comments that target groups of interest for secondary analyses. These analyses include evaluating the impacts, such as adverse health outcomes, of social media targeting specific communities (Nguyen et al., 2021). Furthermore, leveraging knowledge of the target identity can better inform interventions or moderation of hateful content (Tekiroglu et al., 2020). Thus, automated approaches to targeted identity prediction could serve these analyses by streamlining the process of labeling new corpora for study.

While some efforts have been made to develop algorithms that predict targeted identity groups, they have largely focused on classifying individual vs. group targets (Zampieri et al., 2019) or implicitly

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 231 - 244

characterizing the target (Waseem and Hovy, 2016). Predictive models capable of identifying a broad range of targeted protected classes have been less studied (Chiril et al., 2022). Hate speech corpora that include the requisite range of targeted identity annotations have been limited until recently, opening the door to a full examination of this problem (Kennedy et al., 2020; Mathew et al., 2020; Kennedy et al., 2022).

In this work, we developed models to predict identity groups targeted by social media comments. Using the *Measuring Hate Speech* (MHS) corpus (Kennedy et al., 2020), we trained neural networks to predict 8 identity group and 12 sub-group targets of hate speech. We demonstrated that these models exhibited good predictive performance, validating them within the MHS corpus and on external datasets. Lastly, we examined model performance on comments with multiple targets, finding that performance depended highly on those targets.

2 Related Work

Hate Speech Detection and Measurement. This work builds on the long line of work investigating automated hate speech detection (Waseem and Hovy, 2016; Waseem, 2016; Davidson et al., 2017; Del Vigna et al., 2017). Currently, the state-of-the-art approaches utilize large-scale transformer models with transfer learning to detect hate speech (Koufakou et al., 2020; Tran et al., 2020). We use similar approaches in this work.

Targeted Identity Detection. Most work investigating the identification of identity targets in hate speech has viewed it as a sub-task of hate speech detection (Waseem et al., 2017). Several works focused on hate speech detection have implicitly considered target identity via labels that contain information about the target of the speech, such as "racism", "sexism", and others (Kwok and Wang, 2013; Waseem and Hovy, 2016; Indurthi et al., 2019; Grimminger and Klinger, 2021). Other work has considered hate speech targets in the context of "single" or "group" targets. Notably, the shared task OffensEval 2019 (Zampieri et al., 2019) included single vs. group target identification, which has been used in subsequent multi-task frameworks (Plaza-del Arco et al., 2021). Lastly, Mossie and Wang (2020) consider the identification of ethnic groups in Ehtiopian social media comments.

Several works have sought to define the notion of "targeting" while providing analysis on what groups are targeted (ElSherief et al., 2018; Silva et al., 2016). These works largely used rules or lexica based approaches for detection. Shvets et al. (2021) explicitly define a "target" and corresponding "aspects", while developing neural networks to extract text matching these concepts in comments.

The creation of corpora that provide labels on targeted identity groups have allowed further analysis of targeted identity prediction (Mathew et al., 2020; Kennedy et al., 2020, 2022). Most relevant to this work is an analysis by Chiril et al. (2022) examining multi-task target identity prediction on a wide range of past corpora. Our study builds on these works by examining the performance on a thorough range of both broad target identity groups and more specific sub-groups.

3 Methods

All code used in this work is available on the `hate_measure` repository[1], which contains a codebase of various models applicable to the MHS dataset, and the `hate_target` repository[2], which contains the code used for the analyses and figures described in this paper. All datasets were obtained as described by their corresponding entries on the Hate Speech Data website (Vidgen and Derczynski, 2020).

3.1 Datasets

We trained and evaluated all models on the Measuring Hate Speech (MHS) corpus created by Kennedy et al. (2020). We performed additional generalization evaluations on two other corpora: the HateCheck Corpus (Röttger et al., 2021) and Gab Hate Corpus (GHS) (Kennedy et al., 2022). We chose to train on the MHS corpus because it was the largest dataset that covered a diverse range of platforms.

Measuring Hate Speech. The MHS corpus was constructed to facilitate the measurement of hate speech with item response theory. It consists of 50,070 hate speech comments obtained from Reddit, Youtube, and Twitter, labeled by 11,143 annotators. Annotations consisted of 10 survey items spanning a theorized spectrum of hatefulness. Additional annotations, of main interest for this work, included the target of the comment. Specifically, annotators were asked *"Is the [comment] directed at or about any individuals or groups based on..."*,

[1] https://github.com/dlab-projects/hate_measure

[2] https://github.com/dlab-projects/hate_target

with the option to select among the following eight identity groups: race/ethnicity, religion, national origin or citizenship status, gender, sexual orientation, age, disability status, political identity; or "none of the above". Annotators could select more than one identity group. We note that the MHS corpus allows target identity annotations to include those that are the subject of supportive speech. Thus, "target" within the scope of this dataset can be understood to mean the identity group a comment speaks to, whether it is hateful or supportive.

For each identity group selected (if any), the annotator was prompted to select identity sub-groups. For example, if the annotator indicated a target based on race/ethnicity, they were asked to specify racial/ethnic sub-group identities, including: Black/African American, Hispanic/Latino, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic White, or an "Other" category with the option to provide written text. As another example, the possible sub-groups for gender identity included Men, Women, Non-binary, Transgender Men, Transgender Women, or Transgender unspecified (along with an "Other" category allowing for annotator specification). See Appendix B for all identity sub-groups.

HateCheck Corpus. The HateCheck Corpus is comprised of a set of functional tests for hate speech detection models. The samples in HateCheck are synthetically constructed to allow diagnostic assessment of model performance. These synthetic expressions generally make apparent who the target is, e.g., "I hate [IDENTITY GROUP]". Thus, they serve as a useful sanity check for validating the performance of a model.

The HateCheck Corpus contains 3,901 comments, of which 3,606 have a labeled target. These targets are specifically labeled as "gay people", "women", "disabled people", "Muslims", "black people", "trans people", and "immigrants". To evaluate generalization performance, we recast these labels as follows: "gay people"→ Sexual Orientation, "women" → Gender Identity, "disabled people" → Disability, "Muslims" → Religion, "black people" → Race, "trans people" → Gender Identity, and "immigrants" → National Origin.

Gab Hate Corpus. The Gab Hate Corpus (GHC) is comprised of 27,665 posts from the social media platform Gab (Kennedy et al., 2022). Using a hierarchical coding typology, The posts were annotated for "the presence of hate-based rhetoric." The corresponding identity group targets include nationality/regionalism, race/ethnicity, gender identity, religious/spiritual identity, sexual orientation, ideology, political identification, and mental/physical health status. We recast the ideology and political identification labels as a single "political ideology" label and map the remaining groups directly onto those of the MHS corpus.

The GHC only includes target identity labels if the comment expressed hate toward those target identities. Since the MHS corpus includes target identity labels for either hateful or supportive speech, we omitted samples in the GHC which lacked target identity labels, resulting in a subcorpus of 7,801 comments. We did this since a model trained on the MHS may predict targets for the GHC that would have no corresponding label, since annotators would not have identified targets if they did not deem the comment hateful.

3.2 Data Preparation

We performed minimal preprocessing on each data sample, including normalizing blank space and replacing URLs, phone numbers, and emails with respective tokens. We then passed each comment through a tokenizer corresponding to the base model architecture being trained.

We formulated the task of predicting targeted identities as a multi-label binary prediction. However, each comment was annotated by more than one annotator. Annotators expressed moderate agreement on identifying the targeted groups, with Krippendorff's alphas ranging from $0.6 - 0.75$ (see Appendix C). We used soft labeling for training, where the proportion of annotators identifying an identity group as a target served as the "label". When calculating evaluation metrics, we only used binary labels by majority voting.

Following Kennedy et al. (2020), we removed annotators according to two quality checks revolving around the infit mean-square statistic (Linacre et al., 2002), and satisfactory identification of target identities. Filtering annotators according to these quality checks resulted in 8,472 annotators remaining, with 39,565 accompanying comments.

3.3 Model Architecture

We tested various pre-trained transformer architectures in predicting the multi-label binary outcome. Specifically, we used Universal Sentence Encoder (Cer et al., 2018), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019) as base models. We

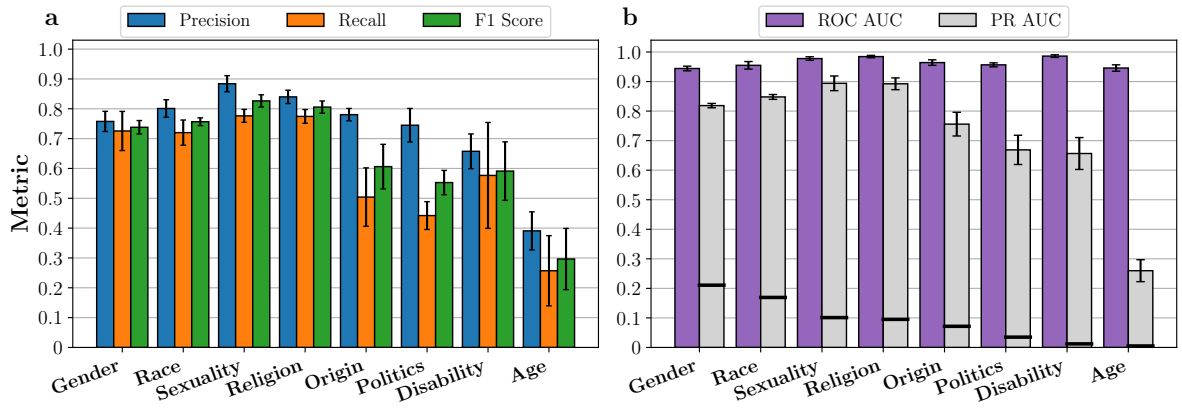

Figure 1: **Transformer models are predictive of target identity groups.** The performance on target group identity prediction across test folds of the MHS corpus as quantified by threshold-dependent and threshold-agnostic metrics. Error bars denote the standard deviation across the test folds. **a.** Precision, recall, and F1 score on test set data according to a 0.5 threshold, for each target group identity. **b.** ROC and PR AUC on test set data. Black lines denote the incidence rate (proportion of positive labels) of the corresponding target identity group. Identity groups are sorted in order of decreasing incidence rate.

stacked a feedforward layer on top of the model embeddings, and then placed M binary output layers, where M is the number of output groups under consideration. We applied dropout to the feedforward layer, with the specific rate chosen as a hyperparameter. We used pre-trained models obtained from HuggingFace (Wolf et al., 2020).

3.4 Training Procedure

We considered a variety of hyperparameter configurations when training models, varying the size of the dense layer, the batch size, and the dropout rate. The full set of configurations is listed in Appendix A. We used a validation set to determine the number of epochs to train on, as described below. We additionally weighted each sample by the square root of the number of annotators. Lastly, we used cross-entropy as the loss function for each output, and used the sum of individual losses as the loss for the entire network.

We performed 5-fold cross validation to train and evaluate models. After shuffling the data across samples, we split the dataset into 5 folds. For each architecture, we trained 5 models, each using 4 folds for training and the remaining fold for evaluation. Each training fold was further split into training and validation sets. We then trained the model using the training set data with early stopping on the validation loss. When validation performance decreased past epoch E, we halted training, and retrained the model on the entire training fold for E epochs. We then evaluated the model performance on the test fold. Model evaluation metrics were reported across the 5 test folds. For out-of-corpus generalization tasks, we applied a model trained on the entire dataset, using the average number of epochs across folds during cross-validation.

3.5 Evaluation Metrics

Since most labels we considered were imbalanced, we evaluated an array of complementary metrics. As is commonly done, we focused on a set of threshold-dependent metrics (precision, recall, F1 score) and threshold-agnostic metrics (ROC AUC and PR AUC) in the main text. We report two additional metrics–the accuracy over chance and log-odds difference–in the Appendix.

We used traditional threshold-dependent metrics capturing false positive/false negative rates, including the *precision*, *recall*, and *F1 score*. We calculated these metrics using predictions at a threshold of 0.5, unless otherwise specified. We supplement the traditional metrics with threshold-agnostic metrics, including the area under the receiver operater characteristic curve (ROC AUC), and the area under the precision-recall curve (PR AUC). Importantly, we use the PR AUC in addition to ROC AUC as it may be more informative in imbalanced datasets (Davis and Goadrich, 2006). We used macro-averaging to summarize a metric across labels. This process consisted of weighting each label's performance metric by their incidence rate when calculating an overall average.

We considered two additional metrics: *accu-*

racy over chance and the *log-odds difference*. For brevity, we describe them here, but report their values in Appendix A. We considered accuracy divided by chance performance in order to confirm that models did in fact generalize beyond that of a naive classifier which could artificially achieve high accuracy in imbalanced settings. In *highly* imbalanced settings (i.e., fewer than 1% of the labels in the positive class), accuracy over chance may not sufficiently capture the performance of a predictive model. This stems from the difficulty in improving performance in highly accurate regimes (e.g., it is more difficult to improve from 99% to 99.5% than 90% to 90.5% accuracy). Thus, we additionally turn to the log-odds difference:

$$\text{LOD} = \log\left(\frac{a}{1-a}\right) - \log\left(\frac{b}{1-b}\right) \quad (1)$$

where a is the test set accuracy and b is the baseline accuracy (e.g., chance). The log-odds difference more effectively weights the difficulty in achieving performance gains when the dataset is heavily imbalanced (e.g., the second term is very large).

4 Results

Our main goal was the multi-label binary prediction of target identity groups. We first trained and evaluated models to predict the targeting of the broad identity groups. We repeated these experiments, but on identity sub-group predictions. We then evaluated the performance of the model on two additional datasets: the HateCheck and Gab Hate Corpora. Lastly, we evaluated the performance of the model on samples which had multiple targets.

4.1 Targeted Identity Group Prediction

We first considered the task of predicting the identity group(s) targeted by a comment. We constructed a multi-label binary prediction task, with the binary outcomes corresponding to gender, race/ethnicity, sexual orientation, religion, national origin, politics, disability, and age (ordered in decreasing incidence rate). We then trained a variety of transformer-based neural networks to predict the targeting of each identity group in parallel. Each model consisted of a base network (pre-trained transformer model) stacked with a dense layer mapping onto the 8 identity groups, with variations on the hyperparameter configuration and data preparation. The full set of experiments and architectures, along with their performance, is listed in Appendix A. For brevity, we show results using a RoBERTa-Large base network with soft labels and training samples weighted by number of annotators (see Methods), which exhibited the best performance of the models we considered.

We found that the model generally excelled at predicting the target of the comment, with performance varying according to the incidence rate of the label. We first evaluated model performance using threshold-dependent metrics such as precision, recall, and the F1 score (Fig. 1a). At a threshold of 0.5, the model achieved F1 scores from $0.7 - 0.85$ for the gender, race, sexual orientation, and religion labels. For national origin, politics, disability, and age, the F1 score decreased. This likely corresponds to the decrease in incidence rate for these labels (Fig. 1b: black lines). Additionally, precision generally exceeded recall, indicating that the model generally suffered from false negatives more often than false positives. This implies that the model could fail to identify comments which targeted identity groups, particularly for the national origin and political ideology labels.

We examined the threshold-agnostic labels–ROC AUC and PR AUC–similarly finding that they indicated high predictive accuracy (Fig. 1b). The ROC AUC values for all identity groups were above 0.90. Meanwhile, PR AUC values were above 0.80 for the gender, race, sexual orientation, and religion labels, above 0.60 for the politics and disability labels, and below 0.30 for age. The performance of the PR AUC roughly tracked with the incidence rate (Fig. 1b), as we might expect. We note that the PR AUC may be a better indicator of performance than the ROC AUC due to the imabalanced nature of the dataset (Davis and Goadrich, 2006). Together, these results demonstrate that the model can simultaneously predict several targeted identity groups. However, this performance suffers on identity groups that are less represented in the dataset (e.g., age and disability).

4.2 Targeted Identity Sub-Group Prediction

We next considered the prediction of specific identity sub-groups. For example, secondary analyses on social media comments may be interested in comments targeting a specific gender identity (e.g., comments targeting women). To this end, we evaluated the performance of a similar task–multi-label binary prediction–but the identity sub-groups. We specifically focus on racial/ethnic iden-

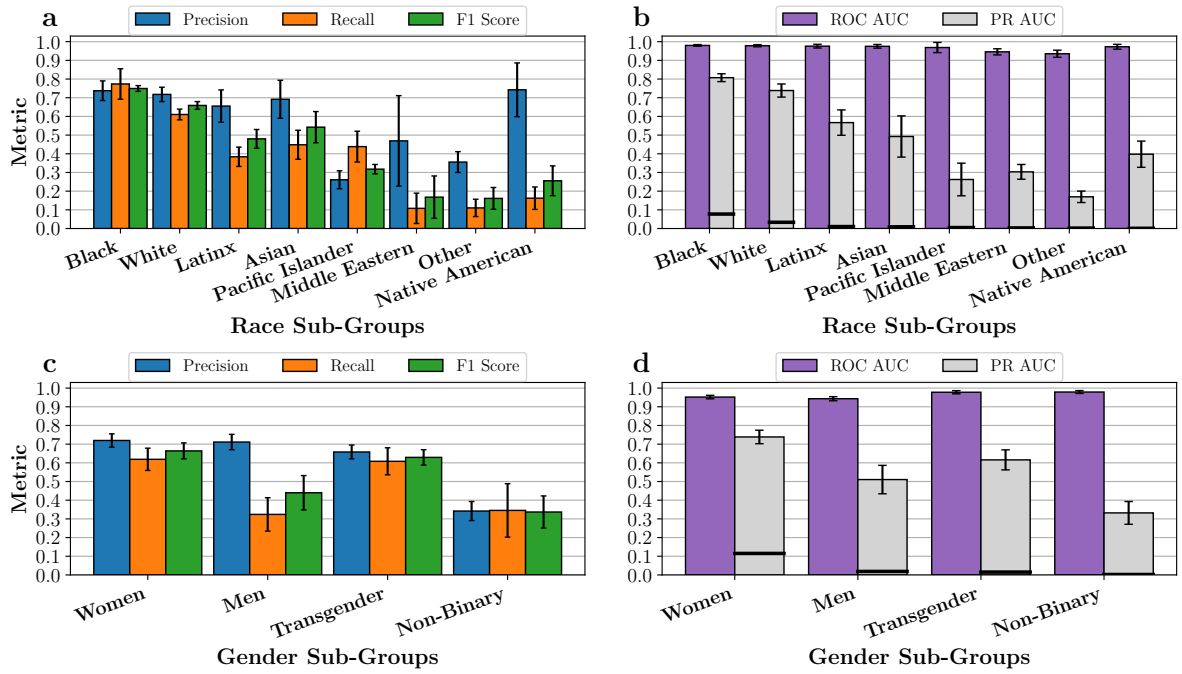

Figure 2: **Model performance on identity sub-groups varies strongly across sub-groups.** The performance on target sub-group identity prediction across test folds of the MHS corpus as quantified by threshold-dependent and threshold-agnostic metrics. **a-b.** Precision, recall, and F1 score on the test set data according to a 0.5 threshold (**a**) and ROC/PR AUCs (**b**) for the racial sub-groups. **c-d.** Same as top row, but for the gender identity groups. Black lines denote the incidence rate (number of positive labels) of the corresponding target identity group. Identity groups are sorted in order of decreasing incidence rate.

tity sub-groups (Black, White, Latinx, Asian, Middle Eastern, Pacific Islander, Native American, or some other group; listed in decreasing order of incidence rate) and gender identity sub-groups (women, men, non-binary; listed in decreasing order of incidence rate) because these groups were the most well-represented in the corpus. Within the gender identity sub-group task, we added an additional transgender label. As in the case of the broader identity groups, we found that the best performing model was a network with a RoBERTa-Large base with soft labels and weighted samples.

We found that the best performing model exhibited high predictive performance on some racial identities (Fig. 2). However, predictive performance was generally lower than that of the group identity prediction. We first evaluated threshold-dependent metrics, finding that the model exhibited the best performance on Black-targeting speech, a median F1 score of 0.72. Similar to the target identity models, precision generally exceeded that of recall, implying the presence of false negatives. These discrepancies were most strongly observed in the racial groups which had the lowest incidence rate, including Middle Eastern, Pacific Islander, Native American, and the Other category (Fig. 2b: black lines). Among the threshold-agnostic metrics, ROC AUC generally indicated superior predictive performance, though this may be a product of label imbalance (Davis and Goadrich, 2006). PR AUC generally tracked with the F1 score (and the incidence rate). A notable exception is Asian identity, which exhibited higher PR AUC than Latinx identity, despite having a lower indicidence rate.

Meanwhile, for the gender sub-groups, we observed worse performance relative to race. The best predictive performance was observed on identifying comments targeting women, with an F1 score of roughly 0.65. Interestingly, we observed substantially better predictive performance in identifying comments targeting transgender people compared to men, despite comparable incidence rates. Overall, we found that the reduced number of samples resulted in decreased predictive performance for many identity sub-groups.

4.3 Models Generalize to External Corpora

Thus far, we have examined model performance on held-out data within the MHS corpus, which consists of comments from Reddit, Twitter, and

HateCheck Corpus

Identity Group	Accuracy (Chance)	F1 Score	ROC AUC	PR AUC
Disability	0.989 (0.869)	0.957	0.996	0.986
Gender	0.978 (0.739)	0.954	0.994	0.990
National Origin	0.986 (0.875)	0.941	0.990	0.972
Race	0.981 (0.871)	0.926	0.990	0.972
Religion	0.984 (0.869)	0.935	0.967	0.951
Sexual Orientation	0.993 (0.852)	0.974	0.991	0.981

Gab Hate Corpus

Identity Group	Accuracy (Chance)	F1 Score	ROC AUC	PR AUC
Disability	0.972 (0.969)	0.237	0.857	0.408
Gender	0.954 (0.927)	0.636	0.939	0.721
National Origin	0.868 (0.846)	0.402	0.821	0.523
Politics	0.788 (0.710)	0.557	0.826	0.667
Race	0.873 (0.781)	0.622	0.880	0.778
Religion	0.924 (0.827)	0.773	0.916	0.763
Sexual Orientation	0.981 (0.954)	0.780	0.948	0.784

Table 1: **Target identity models generalize to out-of-corpus, out-of-platform comments.** The test performance of the target identity model (specifically, the model corresponding to Fig. 1) on the HateCheck (top table) and Gab Hate Corpus (bottom table). The labels provided by each corpus were reassigned to align with the model's outputs (see Methods). Model predictions for identity groups without a corresponding label (age and political affiliation for HateCheck; age for GHC) were discarded. F1 score is calculated with a threshold of 0.5.

YouTube. However, past work has found that hate speech models exhibit a drop in performance on external corpora, particularly when those corpora are sourced from other platforms (Koufakou et al., 2020; Arango et al., 2019). Therefore, we sought to assess out-of-corpus/platform performance of the trained model by evaluating it on two corpora: the HateCheck corpus and Gab Hate Corpus (GHC).

We first considered the HateCheck corpus because it served as a sanity check for model validation. The HateCheck corpus consists of functional tests for hate speech, which often clearly make apparent the targeted identity group (Röttger et al., 2021). Due to the relatively simple syntactic structure, we should expect a trained model to perform well at identifying targeted identities. We relabeled the HateCheck identity groups to align with the trained model, matching to 6 of its 8 identity groups (see Methods). We applied our model to all samples in the corpus and evaluated the performance.

We found that the model exhibited superior predictive performance on the HateCheck corpus (Table 1: top). We obtained accuracies ranging from $0.97 - 0.99$ for each identity group, greatly exceeding that of chance, which ranged from $0.7 - 0.86$. At a threshold of 0.5, F1 scores were all above 0.90. Meanwhile, AUC scores were well above 0.95 for all identity groups, implying tight control of false positives and false negatives.

We supplemented the above generalization check with the Gab Hate Corpus (GHC), consisting of comments extracted from the social media platform Gab (Kennedy et al., 2022). The GHC covers a wide range of target group identities that match closely with those of the MHS corpus. Furthermore, it presents a useful test case to evaluate the extent to which the target identity model generalizes to a new distribution of comments. We applied our model to the subset of comments on which the annotators specified a hateful target (see Methods).

We found that the model generally performed well on the GHC, but exhibited a slight drop in predictive performance relative to the MHS corpus (Table 1: bottom). The model achieved accuracies ranging from $0.78 - 0.98$, well above chance. The model exhibited wide ranging F1 scores, with poor or average performance on the disability, national origin, and political affiliation groups. The ROC AUC and PR AUC scores similarly suggested good predictive performance, but were lower than those on the MHS corpus. Tracking with incidence rate, the model exhibited the best performance on the gender, race, religion, and sexual orientation categories. Overall, these results demonstrate that the

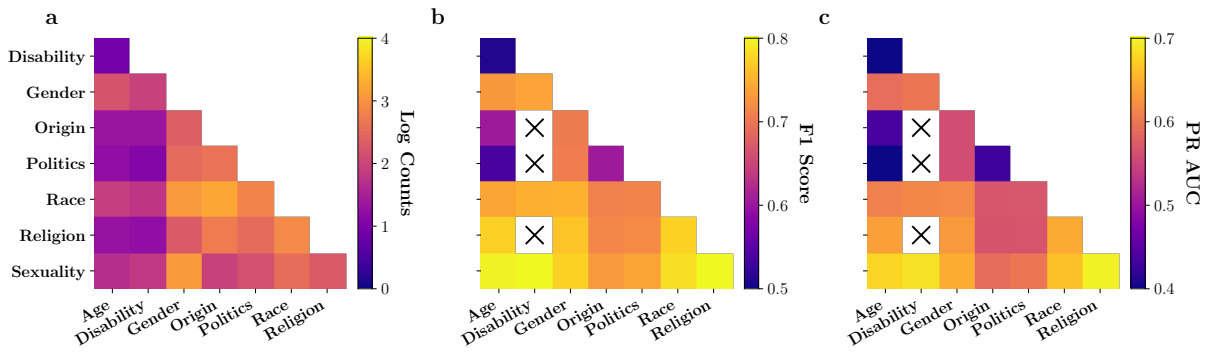

Figure 3: **Models exhibit diverse performance on multi-target samples. a.** The log-count of samples for each pair of identity groups in the MHS corpus. **b.** The macro-F1 score evaluated on sub-corpora containing samples in which each pair of identity groups was targeted (according to annotators) or predicted to be targeted by the classifier. **c.** The PR AUC on the same sub-corpora, across identity group pairs.

predictive models generalize fairly well to novel, out-of-platform data.

4.4 Model Performance on Multiple Targets

Hate speech can target multiple identity groups, either referencing them as separate targets (e.g., referencing a Black person and woman separately) or as a single, intersectional target (e.g. referencing a Black woman, a single subject with racial and gender identity components). We sought to examine how well the classifier performed in scenarios where two identities were targeted in the same comment, either by annotation or prediction.

We first examined the number of comments for each pair of target identity groups in the corpus. We assigned binary labels based on annotator majority voting for each target. Then, for each pair of identity groups, we calculated the number of comments which targeted both identity groups. The distribution of log-counts for each pair of identity groups is shown in Figure 3a. These counts generally aligned with the number of samples for each identity group. For example, (gender, race), the two largest identity groups in the corpus, had among the highest log-counts. However, the relationship between the identity groups also played a role in the observed counts. For example, (race/ethnicity, national origin) and (gender identity, sexual orientation) were the two combinations with the largest number of samples. This likely stems from the topic overlap within each pair.

We might expect a classifier to perform well on identity group pairs with a large number of samples. The classifier could, however, produce errors on these pairs by mistaking one identity group for another. Furthermore, the classifier may predict multiple targets when only one target is present. In order to evaluate the performance of the model in these settings, we consolidated a sub-corpus of comments for which (i) annotators identified two targeted identity groups or (ii) the classifier identified two targeted identity groups. Thus, the sub-corpus could contain either false negatives (classifier failed to predict both identity groups) or false positives (classifier mistakenly identified multiple identity groups). For each pair of identity groups, we calculated the average F1 score and PR AUC across the pair of labels (weighted by incidence rate). We note that we could only calculate these metrics when the classifier exhibited some false positives. If this did not occur, the F1 score and PR AUC would be undefined. We denote these rare instances with an X in Figure 3.

We examined the distribution of the F1 score and PR AUC across the pairs of identity groups (Fig. 3b-c). We found that, generally, the model exhibited worse performance on identity pairs which had the least number of samples, such as (age, disability) and (age, politics). On the other hand, the model generally performed well in cases where there were an abundance of samples, such as (race, gender). However, we observed other interesting relationships. For example, the model exhibited the best performance for identity pairs that were less related to each other, such as (age, sexual orientation), despite these pairs having lower counts. Notably, (origin, politics) exhibited markedly lower predictive performance, despite having more samples than other pairs. Together, these results highlight that performance on samples with multiple identity groups is modulated by the identity group pair under consideration.

5 Discussion

We have demonstrated that transformer-based neural network models can achieve good predictive performance on classifying multiple targeted identity groups or sub-groups simultaneously. We additionally validated the models on out-of-corpus data, finding that the results indicated some degree of generalizability. These results largely serve to benchmark this task for future studies, but also raise additional questions on the definition and conceptual framing of "targeting" in hate speech corpora.

We evaluated the performance of the model on multiple targets. However, the survey question prompting for identity targets did not distinguish between a *single* target with multiple identities, or *multiple* distinct targets. For example, a secondary analysis may be interested in comments that target Black women (at the intersection of racial and gender identity sub-groups), which are distinct from comments that separately target a Black person and a woman, but would be indistinguishable under the labeling scheme. The distinction is important, as the former setting corresponds to intersectional identity (Crenshaw, 2018), on which datasets and machine learning algorithms have been demonstrated to exhibit biased coverage or performance (Kim et al., 2020). Thus, the development of new labeling instruments that ask annotators to make the distinction between intersectional and multiple targets is of interest for future work. For example, Fortuna et al. (2019) developed a hierarchical labeling scheme which allowed for the the identification of intersectional targets in a Portugese dataset.

In this work, we considered multi-label networks designed to simultaneously predict either identity groups or sub-groups. However, constructing networks that can simultaneously predict multiple *sets* of sub-groups is of interest, particularly for identifying intersectional targets in social media content. This can be viewed as *multi-task* problem, which may require adjustment to network architectures in order to achieve desirable performance (Crawshaw, 2020; Talat et al., 2018). The development of multi-task networks with identity group specific sub-networks is of interest for future work (Plaza-del Arco et al., 2021). Such networks could, for example, contain sub-networks predicting racial identity sub-groups, gender identity sub-groups, and others, in parallel.

We relied on synthesizing annotator responses into a single label for each comment, while incorporating some knowledge of their disagreement. This approach generally falls in line with the weak perspectivist approach in predictive computing (Basile et al., 2021). However, annotator disagreement on the identity group targets (Appendix C) indicates that there is some subjectivity in identifying targeted groups. Data perspectivist approaches more strongly incorporating different annotator responses are a viable path forward (Basile et al., 2021; Sudre et al., 2019; Uma et al., 2020). At the same time, continued improvement in labeling instruments could further ameliorate these issues. For example, instruments that allow annotators to explain their reasoning in a structured fashion could shed light on why annotator disagreement is present. Qualitative examination of comments could support additional theorization of the the concept of "targeting". In this vein, following Kennedy et al. (2020), it may be possible to develop a measurement scale for "targeting" to facilitate item response theory approaches on this task.

Extensions to this work could facilitate parsing of the sentence to better elucidate the manner in which hateful comments refer to targets. For example, Shvets (2021) develop extraction networks to identify the text corresponding to both the "target" of a comment and its "aspect", or the characteristic attributed to the target. Such work could facilitate additional qualitative examination of comments.

While hate speech is understood to "target" a person or group based on a characteristic, the notion of "targeting" is slightly different across datasets. For example, we used "target" to mean the identity group that a comment is directed toward, whether the comment exhibited positive or negative valence. This was framed in the context of a measurement scale spanning supportive and hateful speech (Kennedy et al., 2020). However, other corpora limit their definition to content that is strictly hateful. These subtle distinctions limit the ability of out-of-corpus validation on datasets. For example, in this context, we could only use a subset of the GHC for generalization, since many comments were deemed not hateful (and thus did not have targeted identity annotations), despite referencing an identity group. Datasets may also reference the manner in which "targeting" occurs, such as calls to violence, usage of profanity, or implicit rhetoric (e.g., sarcasm or irony). Further work is needed to standardize these definitions to better inform the curation of future corpora.

Acknowledgements

We thank members of the D-Lab for useful feedback and discussions.

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Kimberlé Crenshaw. 2018. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. In *Feminist legal theory*, pages 57–80. Routledge.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: incorporating lexical features with bert for the detection of abusive language. In *Fourth Workshop on Online Abuse and Harms*, pages 34–43. Association for Computational Linguistics.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

John M Linacre et al. 2002. What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions*, 16(2):878.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Thu T Nguyen, Shaniece Criss, Eli K Michaels, Rebekah I Cross, Jackson S Michaels, Pallavi Dwivedi, Dina Huang, Erica Hsu, Krishay Mukhija, Leah H Nguyen, et al. 2021. Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. *SSM-population health*, 15:100922.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Andrew Sellars. 2016. Defining hate speech. *Berkman Klein Center Research Publication*, 2016(20):16–48.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

Anna Shvets. 2021. System description for the CommonGen task with the POINTER model. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 161–165, Online. Association for Computational Linguistics.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media*.

Carole H Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, et al. 2019. Let's agree to disagree: Learning highly debatable multirater labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Alexander Tsesis. 2002. *Destructive messages: How hate speech paves the way for harmful social movements*, volume 27. NYU Press.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Zeerak Waseem. 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 Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Richard Ashby Wilson. 2017. *Incitement on trial: Prosecuting international speech crimes*. Cambridge University Press.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Extended Experiment Results

Base Model	Hyperparams	Acc/Chance	LOD	AUC ROC	PR ROC	F1 Score
USE V4	Binary Labels H256 B32 D0.1	1.062	0.941	0.949	0.498	0.428
USE V4	Soft Labels H256 B128 D0.1	1.130	1.131	0.938	0.607	0.529
DistilBERT	Binary Labels H256 B64 D0.1	1.135	1.179	0.942	0.648	0.597
DistilBERT	Binary Labels H128 B64 D0.1	1.136	1.203	0.940	0.650	0.584
BERT Base	Binary Labels H128 B32 D0.1	1.137	1.215	0.942	0.667	0.610
BERT Base	Soft Labels H128 B32 D0.1	1.138	1.243	0.952	0.681	0.594
BERT Base	Soft Labels Weighted Samples H128 B32 D0.1	1.139	1.259	0.952	0.682	0.597
RoBERTa Base	Binary Labels H128 B32 D0.1	1.137	1.202	0.947	0.660	0.609
RoBERTa Base	Soft Labels Weighted Samples H128 B32 D0.1	1.139	1.231	0.952	0.673	0.593
RoBERTa Large	Soft Labels Weighted Samples H256 B8 D0.05	1.164	1.343	0.964	0.724	0.647

Table 2: Full experimental results. LOD denotes "log-odds difference". USE denotes "Universal Sentence Encoder". "H" denotes the size of the hidden layer. "B" denotes batch size. "D" denotes dropout rate. Metrics are calculated by averaging across identity groups.

B Annotator Identity Groups and Sub-Groups

Identity Group	Identity Subgroups
Race or ethnicity	Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white
Religion	Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims
National origin	A specific country, immigrant, migrant worker, undocumented person
Gender identity	Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)
Sexual orientation	Bisexual, gay, lesbian, heterosexual
Age	Children (0 - 12 years old), adolescents / teenagers (13 - 17), young adults / adults (18 - 39), middle-aged (40 - 64), seniors (65 or older)
Disability status	People with physical disabilities (e.g., use of wheelchair), people with cognitive disorders (e.g., autism) or learning disabilities (e.g., Down syndrome), people with mental health problems (e.g., depression, addiction), visually impaired people, hearing impaired people, no specific disability

Table 3: Identity group and corresponding subgroups annotators were asked to identify as targets of comments.

C Annotator Agreement on Targeted Identity Groups

Identity Group	Krippendorff's Alpha
Age	0.341
Disability	0.744
Gender Identity	0.712
National Origin	0.571
Race	0.672
Religion	0.797
Sexual Orientation	0.718

Table 4: Annotator agreement on target identity group labels, calculated across samples with Krippendorff's alpha.