HateModerate: Testing Hate Speech Detectors against Content Moderation Policies

Warning: this paper discusses and contains content that can be offensive or upsetting.

Anonymous ACL submission

Abstract

To protect users from massive hateful content, existing works studied automated hate speech detection. Despite the existing efforts, one question remains: do automated hate speech detectors conform to social media content policies? A platform's content policies are a checklist of content moderated by the social media platform. Because content moderation rules are often uniquely defined, existing hate speech datasets cannot directly answer this question.

This work seeks to answer this question by creating HateModerate, a dataset for testing the behaviors of automated content moderators against content policies. First, we engage 28 annotators and GPT in a six-step annotation process, resulting in a list of hateful and non-hateful test suites matching each of Facebook's 41 hate speech policies. Second, we test the performance of state-of-the-art hate speech detectors against HateModerate, revealing substantial failures these models have in their conformity to the policies. Third, using HateModerate, we augment the training data of a top-downloaded hate detector on HuggingFace. We observe significant improvement in the models' conformity to content policies while having the comparable scores on the original test data. Our dataset and code can be found in the attachment.

Hate Speech Community Standards Guidelines

Tier 1: Dehumanizing Speech
- Compare the protected groups as animals that are perceived as inferior (*including but not limited to: apes, pigs*)

Tier 2: Contempt Despise
- Expressions of hate (including but not limited to: despise, hate)

Additional Enforcement: Change Sexual
- Content explicitly providing or offering to provide products or services that aim to change people's sexual orientation or gender identity.

Figure 1: Examples of community standards guidelines for hate speech (Facebook, 2022)

1 Introduction

Social media platforms such as Facebook, Reddit, and Twitter/X have facilitated users to exchange information, but they also expose users to undesirable content, including hateful speech, misinformation, graphic violence, and pornography. To protect users from a massive amount of hateful content, existing work has been vigorously investigating new NLP approaches and providing new resources and open-source tools for studying hate speech detection (Talat and Hovy, 2016; Davidson et al., 2017; Vidgen et al., 2021; Mathew et al., 2021; Hartvigsen et al., 2022; Antypas and Camacho-Collados, 2023). Meanwhile, platforms also invested and achieved great success in building content moderation tools (Facebook, 2023; OpenAI, 2023b), e.g., Facebook's automatic content moderator detected 95% unwanted content before it is seen by a user (Facebook, 2023).

Despite the existing work on hate speech, there remains an important question that is not well addressed: Do hate speech detectors' behaviors conform to platforms' content policies? Content policies are platform-specified rules on what content it moderates. For example, as of Nov 2022, Facebook specifies 41 community standards guidelines for moderating hate speech (Facebook, 2022); Figure 1 shows 3 examples of Facebook's guidelines. The content policies serve as a "contract" between users and the platform; without conforming to the policies, the decision on automated content moderators may be surprising to users, undermining the transparency and accountability of the moderation system. Such trustworthiness issues have led to incidents such as Reddit blackouts, which prevent users from accessing the contents normally (Matias, 2016). Meanwhile, the answer to this question cannot be directly addressed using existing hate speech datasets. The reason is that many platforms have unique moderation rules, e.g., Facebook moderates advertisements on homosexual therapies. Our investigation shows that these custom rules are not well represented in existing hate speech datasets, causing an underestimation of the models' failures in conforming to these rules.

To assess the conformity of automated content moderators to content policies, this paper proposes a dataset called HateModerate, which consists of 7.6k hateful and non-hateful examples for the 41 community standards guidelines on Facebook. Among the published moderation rules from existing work (Banko et al., 2020; Facebook, 2022; Röttger et al., 2021), we opt for Facebook's community standards guidelines for hate speech (Facebook, 2022) as previous work shows it is the most comprehensive among all platforms (Jiang et al., 2020) and it has good clarity.

HateModerate is constructed using the six-step process illustrated in Figure 2. First, we recruit a group of 28 graduate students as the annotators. A part of these students manually search for hateful examples from existing datasets matching each policy. Second, since some guidelines contain too few matched examples, we augment these guidelines by generating hateful examples with the GPT engine. Third, to ensure that the searched and generated examples indeed match the criteria, 16 additional annotators manually verify each hateful example. Fourth, after the hateful examples are collected, for each guideline, we retrieve difficult non-hateful examples from existing datasets that closely resemble the hateful examples to help detection the model failures. Fifth, similarly, we augment guidelines with GPT-generated non-hateful examples. Sixth, 4 additional annotators manually verify each non-hateful examples. The average agreement rate (Krippendorf's alpha) on the match/unmatch of hateful and non-hateful examples are 0.521 and 0.809.

After constructing HateModerate, we examine state-of-the-art hate speech detectors against each policy using the dataset. More specifically, we examine the following models: Google's Perspective API (Google, 2023b), OpenAI's Moderation API (OpenAI, 2023a), Facebook's RoBERTa model (Facebook, 2021) and Cardiff NLP's RoBERTa model (Antypas and Camacho-Collados, 2023). We make the following observations. First, all models prioritize more severe policies (e.g., violence) compared to less severe policies (e.g., stereotyping); second, the OpenAI model conforms the best to the content policies; third, besides OpenAI, models generally have high failure rates for non-hateful examples. After observing the model failures, we further seek answers on how to improve the models' conformity to policies. By adding HateModerate to the training dataset of a top-downloaded model on HuggingFace, the model's performance on HateModerate and HateCheck (Röttger et al., 2021) is significantly improved while the performance on the original test set remains comparable. This result highlights the importance of our dataset in improving the model conformity to content policies.

2 Background and Related Work

2.1 Hate Speech Detection

Construction of Hate Speech Datasets. Automatically detecting hateful speech online is a challenging problem in natural language processing. In recent years, hate speech detection benefits from the advancement of machine learning and NLP techniques (He et al., 2024; OpenAI, 2023b); nevertheless, previous work argues that the datasets play a more important role than the model architecture in hate detection (Gröndahl et al., 2018). Existing work has contributed to many public datasets for hate speech detection (Talat and Hovy, 2016; Davidson et al., 2017; Vidgen et al., 2021; Mathew et al., 2021; Hartvigsen et al., 2022). Since hate speech constitutes approximately 1% of all online speech (Sachdeva et al., 2022), previous work leverage different sampling techniques to improve the efficiency of labeling. For example, by using pre-defined keywords and Twitter hashtags (Davidson et al., 2017; He et al., 2021; Talat and Hovy, 2016; Golbeck et al., 2017). However, hard filtering based on keywords may lead to low coverage issues (Sachdeva et al., 2022). Alternatively, previous work employed information retrieval (Rahman et al., 2021) and classification to create a soft filter (Sachdeva et al., 2022). Our work does not have the class imbalance problem as we reuse the existing hate speech datasets. We further improve the coverage of the dataset with GPT-generated examples.

The Taxonomy for Hate Speech Detection. A taxonomy defines what content is considered hateful. A taxonomy with detailed guidelines can help non-expert annotators better understand the labeling goal. The guidelines contain a checklist of

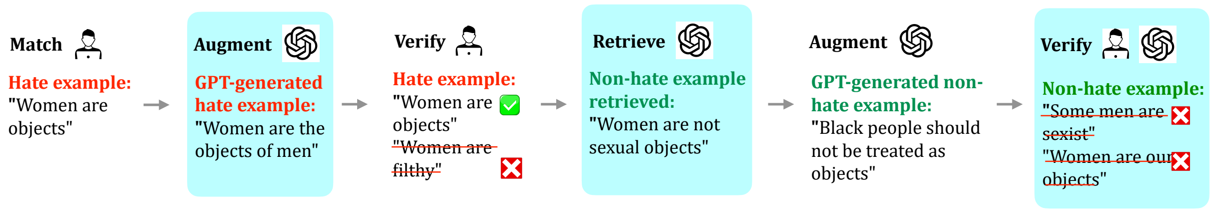

Figure 2: The workflow of data collection for Guideline 10 (Tier 1, Certain objects).

descriptions of the hateful and non-hateful content (Talat and Hovy, 2016; Sachdeva et al., 2022; ElSherief et al., 2021); some previous work further provides codebooks containing more detailed instructions on what is not considered as hateful for each guideline (Golbeck et al., 2017; Vidgen et al., 2021). Banko et al. (Banko et al., 2020) introduce a unified taxonomy of harmful content, including *sexual aggression, doxxing, misinformation* and *hate speech*. Our annotators are provided with Facebook's 41 community standards guidelines. These guidelines contain fine-grained categories (e.g., subcategories of dehumanization) of hate speech as well as new categories that are not well covered in existing datasets (e.g., advertisements of homosexual therapies).

2.2 Policies for Content Moderation

Regulations of Governments/Councils. Online content moderation is subject to policies and regulations of the governments (Congress, 1996; Union, 2022). Zufall et al. (Zufall et al., 2022) constructs a "punishable" hate speech dataset in German based on the German Criminal Code and a legal decision framework. Chiril et al. (Chiril et al., 2021) study gender bias based on the definition by the French High Council on Gender Equality.

Social Media Content Policies. Although platforms have the right to decide what content to moderate (Congress, 1996), users show concerns over the consistency and transparency of the moderation decisions (Matias, 2016). To improve the transparency of moderation, many major platforms released their content policies (Facebook, 2022; Twitter, 2023; Instagram, 2023; Pinterest, 2023; Reddit, 2020), which serve as a "contract" between the user and the moderation system. The policies are based on what value is preserved by the platform, which vary across platforms, e.g., Gab allows more elitism speeches than Twitter (Zhou et al., 2019). Jiang et al. (Jiang et al., 2020) conduct a comparative study of the existing community standards guidelines across platforms; their study suggests that Facebook's guidelines are the most comprehensive ones above all.

Facebook's Community Standards Guidelines. As of Nov 2022, Facebook provides a list of 41 community standards guidelines for hate speech moderation (Facebook, 2022). Figure 1 shows three examples of Facebook's hate speech guidelines, and Table 7 shows the complete list. Facebook's guidelines are organized into four tiers based on the content severity (Facebook, 2022): Tier 1 includes the most offensive content, e.g., dehumanization and violence towards protected groups; Tier 2, Tier 3, and Tier 4 (the additional enforcement) are less severe, e.g., stereotyping and contempt towards protected groups. In this work, we leverage Facebook's community standards guidelines for constructing our dataset.

2.3 Behavioral/Capability Tests of NLP Models

HateModerate provides fine-grained failure rate estimation for each content policy. To this end, it can be seen as a dataset for *capability tests* (Ribeiro et al., 2020; Röttger et al., 2021; Yang et al., 2022). The traditional held-out tests may overestimate the model performance when the model has bias (Poliak et al., 2018). To alleviate this issue, Riberio et al. (Ribeiro et al., 2020) propose to construct a checklist of out-of-domain test suites for each capability the model should have. In particular, HateCheck (Röttger et al., 2021) provides a list of 29 test suites for hateful and non-hateful capabilities, e.g., "*We are a group of [PROTECTED GROUP].*" is a non-hateful suite. However, most of the test suites of HateCheck focus on defining hate speeches with *syntactic* structures, and HateCheck's rules suffer from a low coverage of the hate speech categories (Section 4.3 of (Röttger et al., 2021)). On the other hand, the test suites of HateModerate focus on semantic categories specified by the guidelines; it also improves the coverage of hateful content compared to HateCheck.

3 Constructing the HateModerate Dataset

In this section, we describe the steps for the construction of HateModerate.

Annotators Recruitment. HateModerate is annotated by 28 graduate students in Computer Science[1]. The annotators are recruited from PhD and Master students at a research lab and students taking a graduate-level NLP course. The annotation process is overseen by two experts in online hate. All participants are compensated with a $20 Amazon e-gift card. The annotator names are anonymized in the dataset. We obtained the annotators' consent, and it was explained to the annotators how the data would be used. More details about the annotator recruitment can be seen in Section 8.

Data Sources. Most of Facebook's community standards guidelines are on general hateful content, e.g., dehumanization. Therefore, existing datasets should already contain examples matching a significant number of guidelines. We thus first try to search for and reuse examples and their hateful/non-hateful labels from existing datasets. By doing so, we reduce the requirement on annotator expertise and avoid introducing additional labeling errors; notably, it is challenging for non-expert annotators to reach a high agreement rate on hateful/non-hateful labels (Mathew et al., 2021). We first instruct the annotators to search in the following datasets: DynaHate (Vidgen et al., 2021), Toxic Spans (Pavlopoulos et al., 2021), Hate Offensive (Davidson et al., 2017), and HateCheck (Röttger et al., 2021). Later the annotators extended the list to include Twitter Hate Speech (AI, 2023), Ethos (Mollas et al., 2020), FRENK (Ljubešić et al., 2019), and COVID Hate and Counter Speech (He et al., 2021). The hateful/non-hateful labels are available in all datasets.

3.1 Collecting Hateful Examples

Manually Searching Matching Hateful Examples. For the first step, we collect the hateful examples matching each guideline. We assign each of Facebook's 41 policies to one of 7 annotators. Annotators are instructed to search for a minimum of 200 hateful examples. Synonyms and regular expressions are employed to enhance the search efficiency. For example, for Guideline 0 (Tier 1, Dehumanize Filth), the annotator uses the regular expression ".*(filth|dirt).*". Multiple annotators also report that they seek help from Google, ChatGPT, and other team members to correctly understand the policy. For example, for Guideline 13 (Tier 1, Hatecrime Victim), the assigned annotator first struggles to find enough examples because the exact word "*hate crime*" rarely appears in any example, after a discussion, the team members suggest him to search for concrete hate crime keywords including *lynching* and *holocaust*. The annotator is able to add more examples as a result.

Data Augmentation for Hateful Examples. After the initial manual search, we find that for some guidelines, the existing datasets do not contain a sufficient number of matching examples, e.g., our annotators only find 3 matching examples for Guideline 38 (Tier 4, Change Sexual). To improve the balance of these guidelines, we further ask the initial annotator to either manually create or use the GPT-3 engine (text-davinci-003) to generate synthetic hateful examples. The prompt we used for the generation is: "*Based on hate speech guideline [GUIDELINE], write a hate speech sentence about [PROTECTED GROUP]*".

Manually Verifying Matched Hateful Examples. In the first step, the hateful examples in each guideline are matched by only one annotator. We need further verification to reduce the errors in the initial annotators' matching. To this end, we involve a total of 16 additional annotators to label whether an initially found example matches (1) or does not match (0) the guideline. We remove an example if at least one annotator labels it as 0. Each example is labeled by 1 to 7 annotators. Following existing work on using demonstrated examples to improve the quality of crowdsourced annotations (Gupta et al., 2023), we provide a few false match examples for each policy, e.g., for Guideline 29 (Tier 2, Curse Sexual), "*They make me so pissed off these immigs!!*" is demonstrated as a false match. The average agreement rate (Krippendorf's alpha) between the annotators is 0.521 (Appendix A.7).

As mentioned above, some guidelines contain few matching examples from existing datasets, we repeatedly perform verification/removal and augmentation until each guideline contains at least 50 valid matching hateful examples.

3.2 Collecting Non-Hateful Examples

Retrieving Difficult Non-Hateful Examples. Since testing with only hateful examples will result in bias (e.g., one model has a low failure rate simply because it sets a low threshold for hate), we further add non-hateful examples to HateModer-

[1]We opt for students labeling rather than Amazon Mechanical Turk labeling since the quality of students' labeling is more manageable, we notice some existing work on hate speech dataset collection also used students labeling (Fanton et al., 2021).

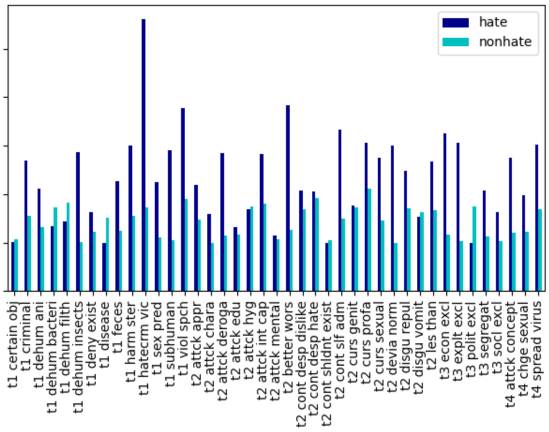

Figure 3: The statistics of examples in each policy in our dataset

ate. To improve the detection of model failures, for each policy, we opt for retrieving more difficult non-hateful examples that are most similar to the hateful examples from the previous stage. The corpus we retrieve from are the non-hateful examples in DynaHate (Vidgen et al., 2021), since DynaHate contains a large number of manually created adversarial non-hateful examples that look similar to hateful examples. The retrieval algorithm follows the state-of-the-art dense retrieval paradigm (Karpukhin et al., 2020). We employ OpenAI's Embedding API (OpenAI, 2023) with the `text-embedding-ada-002` model to obtain the vectors. For each policy, we rank every non-hateful example in DynaHate by its average cosine similarity with the existing hateful examples and keep the top-100 non-hateful examples in this step.

Data Augmentation for Non-Hateful Examples. Similar to hateful examples, DynaHate does not contain enough non-hateful examples matching some guidelines. We thus also perform data augmentation for non-hateful examples. First, we use GPT-3 (text-davinci-003) to generate non-hateful examples using the following prompt: "*Based on hate speech guideline [GUIDELINE], write a sentence about [PROTECTED GROUP] with [NON-HATE TYPE]. Examples: [EXAMPLES].*". In particular, the data augmentation re-balances the non-hateful type (explained below), i.e., supporting, counter-hate, neutral, and offensive speech against non-protected groups. For offensive speech against non-protected groups, it is difficult for GPT-3 to generate matching examples, so we manually create the non-hateful examples.

Verifying Non-Hateful Examples. Similarly, the retrieved and augmented non-hateful examples may not closely match the guideline. For example, for Guideline 10 (Tier 1, Certain objects) on dehumanizing speech as objects, one top-retrieved non-hateful example is: "*Some men are sexist*" which is unrelated to the guideline. For each example, we further involve four annotators to provide labels on whether one example is related to the guideline (1) or not (0). Each example receives 2 labels. We remove an example if at least one annotator labels it as 0. The average agreement rate (Krippendorf's alpha) between the annotators is 0.809 (Appendix A.7).

We further perform the following classification step for the non-hateful examples. For each non-hateful example, we employ GPT-4 and 1 annotator's verification to classify it into five classes[2]: supporting, counter-hate, neutral, offensive speech against non-protected groups, and hateful speech with the wrong label. The first three classes are based on the definition of non-hateful speeches in previous work (Sachdeva et al., 2022), and we identify the 4th class during labeling. The full descriptions of the five classes can be found in Appendix A.2. This classification step allows us to remove the hateful examples wrongly labeled as non-hateful (about 3.6%) and to re-balance the four non-hateful types in the data augmentation.

3.3 Dataset Statistics

In our final HateModerate dataset, we compile 7,704 examples: 4,796 hateful (4,535 unique ones) and 2,908 non-hateful (2,264 unique ones). Some instances are duplicated because a single sentence can fall under multiple guidelines simultaneously. The majority of examples come from DynaHate (5,174), followed by GPT (1,385), HateCheck (457), manual (270), Toxic Span (102), COVID hate (152), Hate Offensive (92), Ethos (12), Twitter Hate (33), Toxigen (8) and FRENK (19).

Figure 3 shows the statistics of HateModerate by policy. Among the 41 policies, the most frequent policy contains 367 examples whereas the least frequent policy contains 103 examples, all policies contain 100 to 250 examples, and the majority policies contain more than 150 examples.

4 Testing Hate Speech Detectors' Conformity with Content Policies

In this section, we employ HateModerate as our evaluation benchmark to assess how hate speech

[2]The prompt we used for GPT-4 classification is: "*Classify the sentence of Question into categories 1-5, number only + [GUIDELINE]+[EXAMPLES]*".

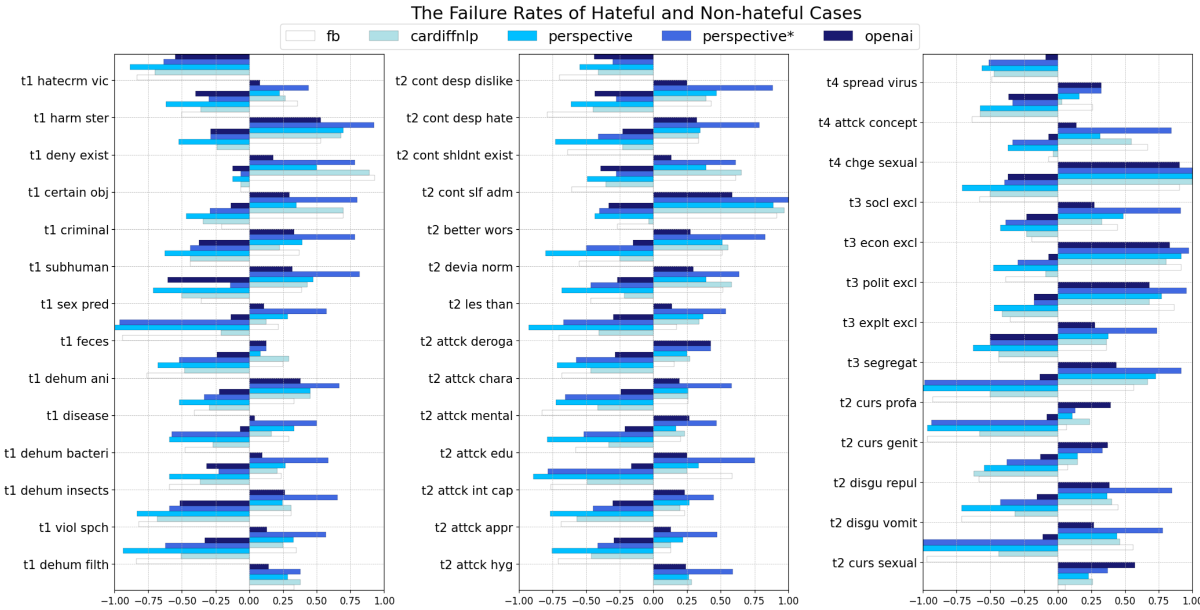

Figure 4: We detect the failure rates for both hateful and non-hateful examples across each of the 41 policies in Facebook's community standards guidelines (Facebook, 2022). Perspective's threshold is 0.5; Perspective*'s threshold is 0.7. For each policy, the bars facing right show the failure rates of hateful examples; the bars facing left show the failure rates of non-hateful examples.

Table 1: The average failure rates of the hateful and non-hateful examples for different tiers of policies, and the average toxicity scores. F: Facebook model, C: Cardiff NLP, P: Perspective with threshold 0.5, P*: Perspective with threshold 0.7, O: OpenAI's API.

T	Failure Rate												Average Toxicity Score									
	Hate						NonHate						Hate					NonHate				
	avg	F	C	P	P*	O	avg	F	C	P	P*	O	avg	F	C	P	O	avg	F	C	P	O
1	.34	.40	.38	.35	.62	**.22**	.47	.52	.39	.65	.43	**.31**	.64	.62	.65	.54	**.74**	.44	.58	.43	.47	**.27**
2	.34	.34	.37	.34	.60	**.30**	.52	.69	.40	.74	.55	**.24**	.62	.64	.62	.55	**.66**	.47	.71	.43	.55	**.21**
3	.59	.63	.57	.66	.90	**.50**	.38	.39	.33	.54	.35	**.27**	.48	.45	.50	.43	**.55**	.33	.42	.31	.38	**.19**
4	.52	.61	.53	.49	.72	**.46**	.36	.40	.36	.50	.39	**.17**	.52	.41	.50	.50	**.68**	.35	.44	.38	.44	**.14**

detectors conform to content policies. We seek answers to the following research questions:

RQ1: How do state-of-the-art hate detectors conform to content policies?

RQ2: What policies do hate speech models conform to the least?

After our initial evaluation, we observe that state-of-the-art models all had different degrees of failures conforming to the content policies. To understand if such failures can be alleviated, we further try fine-tuning existing models with HateModerate. We ask the following research question:

RQ3: Does adding HateModerate to the training data helpimprove a model's conformity to content policies?

4.1 Experiment Setup

Hate Speech Models Evaluated. To answer RQ1-RQ3, we evaluate state-of-the-art models from both industry API endpoints and open-source hate speech detection models. For industry APIs, we choose Google's Perspective API (Google, 2023b) and OpenAI's Moderation API (OpenAI, 2023a; Markov et al., 2023), which are frequently used in downstream detection tasks (Taori et al., 2023; Google, 2023a); for open-source models, we choose Cardiff NLP's fine-tuned RoBERTa model (Antypas and Camacho-Collados, 2023) and Facebook's Fine-Tuned RoBERTa model (Facebook, 2021) which rank top-2 and top-1 among the most downloaded hate models on HuggingFace. The full details of the models can be found in Appendix A.3.

Train/Test Split and Avoiding Data Contamination. To answer RQ3, we reserve 50% of HateModerate for fine-tuning in Section 4.3 by random sampling and use the other half for testing. One issue with evaluating the above models is that their training data may overlap with HateModerate testing data, causing unfair comparisons between models. To minimize the impact of the potential data contamination, for the testing fold, we keep only newly created datasets that are not in the training data of any models. The full details of the excluded data can be found in Appendix A.5.

Evaluation Metric. Following previous work on capability testing (Röttger et al., 2021; Ribeiro et al., 2020), we report the average failure rate of the hateful and non-hateful examples in each policy. If the hateful failure rate is high, it indicates the model cannot effectively detect this category of hate speech; if the non-hateful failure rate is high, it indicates the model cannot effectively recognize non-hateful speeches for that category.

4.2 Evaluating Model Failures using HateModerate

In this section, we seek answers to RQ1 and RQ2. We report the failure rates of each policy in Figure 4. In addition, we report the average failure rate and the average toxicity score of each model in each tier in Table 1. From Figure 4 and Table 1 we can make the following observations.

4.2.1 Comparison of Model Failures of different Policies

From Figure 4 and Table 1, we can observe a contrastive trend between higher tier policies (i.e., more severe hate speech) and lower tier policies. In general, for higher tier policies, the failure rates for hateful examples are lower; whereas the failure rates for non-hateful examples are higher. These results imply that, the less severe speeches specified by lower tier policies are not prioritized by both the industry APIs and the open-source models.

Case Studies of Individual Policies with High Failure Rates. To confirm the validity of our observation, we further conduct case studies on individual policies with the highest failure rates of hateful examples. We identify that most models fail to detect hate speech specified by the following policies: *Certain Objects, change sexual, economic exclusion* and *self admission to intolerance of protected characteristics*. For example, most models fail to detect "*MAKE WOMEN PROPERTY AGAIN!!!!*". We describe these failures in details in Appendix A.1.

4.2.2 Comparison of Different Models' Failures

By comparing the failure rates over all examples (hateful and nonhateful, all tiers) in Table 1, we observe that OpenAI's model exhibits the lowest average failure rate (avg: 0.29, std: 0.17), followed by Perspective (avg: 0.38, std: 0.19). CardiffNLP (avg: 0.40, std: 0.22) and Facebook's RoBERTa (avg: 0.40, std: 0.23) perform less well.

Besides OpenAI, most of the models exhibit high failure rates in non-hateful examples. Perspective with 0.5 threshold performs the worst in non-hateful examples. We further report the failure rate of Perspective with 0.7 threshold in Table 1. We can observe a trade-off between good failure rates in the hateful and non-hateful examples of the two thresholds.

Bias in Toxicity Scoring. In Table 1, we report the average toxicity scores of each model for different tiers of policies, i.e., the probability for the model to predict the hateful class. We can see that while different models have similar toxicity scores for the hateful examples, the scores for non-hateful examples are different. Essentially, Perspective and Facebook's RoBERTa tends to assign higher toxicity for both hateful and non-hateful examples.

Finding Summary of RQ1 and RQ2. ① All models prioritize more severe policies over less severe policies; ② The OpenAI model has the best performance overall, Perspective generally scores sentences with higher toxicity scores, thus a threshold higher than 0.5 is desirable; ③ The models are generally bad at detecting difficult non-hateful examples except for OpenAI (a more detailed analysis can be found in Appendix A.8).

4.3 Mitigating Model Failures with Fine-Tuning HateModerate

In this section, we seek the answer to RQ3. We do so by comparing the failure rates of the following models in Table 2: ① **CardiffNLP**: RoBERTa-base fine-tuned using all the available training data for the CardiffNLP model (Antypas and Camacho-Collados, 2023)[3]; ② **+HM**: RoBERTa-base fine-tuned using CardiffNLP's training data + HateModerate's reserved training data; ③ **+HM***: same as **+HM** but downsample the hateful examples so the hateful and non-hateful examples are balanced; ④ **OpenAI**: The failure rate of the OpenAI API. For the 9 training datasets of the CardiffNLP model, we use the same train/test split as the original datasets[4]. The hyperparameters and more details of fine-tuning can be found in Appendix A.6.

[3] We are only able to access 9 out of the 13 training datasets of the CardiffNLP model. The full details of 9 datasets can be found in Appendix A.4.

[4] Among all 9 datasets, the train/test split is available in only 3 datasets, which we use as the test sets in Table 2. We use all remaining data for the train.

Table 2: The failure rates of fine-tuning with the CardiffNLP data before and after adding HateModerate. Significant results are denoted with †.

FailureRate	Fine-tuned RoBERTa on			
	CardiffNLP	+ HM	+ HM*	OpenAI
HateCheck (Röttger et al., 2021)				
Hate	.442	.185†	.297†	**.008**
Non-hate	.205	.229†	.205	**.016**
Overall	.365	.199†	.235†	**.011**
HateModerate Test				
Hate	.454	**.222**†	.281†	.369
Non-hate	.409	.338†	**.301**†	.351
Overall	.423	**.275**†	.295†	.365
CardiffNLP Test Sets:				
hatEval (Basile et al., 2019)				
Hate	.084	.075	**.061**†	.754
Non-hate	**.776**	.781	.780	.080
Overall	.485	.485	.478†	**.363**
HTPO (Grimminger and Klinger, 2021)				
Hate	.526	.661†	**.525**	.949
Non-hate	.043	.037	.041	**.006**
Overall	.090	.090†	**.089**	.098
HateXplain (Mathew et al., 2021)				
Hate	**.157**	.159	.168	.351
Non-hate	.315	.262†	.266†	**.223**
Overall	.221	**.201** †	.208†	.299

Results of Fine-Tuning. In Table 2, we compare the failure rates on the following test collections: ① The testing fold of HateModerate; ② The 3 testing datasets of CardiffNLP; ③ HateCheck (Röttger et al., 2021), a dataset for independent out-of-domain capability tests of hate speech. We conduct the paired t-test between **+HM** vs **CardiffNLP** and **+HM*** vs **CardiffNLP**. In the **+HM** and **+HM*** columns, we denote the significant results (p-value < 0.05) using †. The details of the t-test results can be found in Table 5 of Appendix A.9. Table 2 reveals that adding HateModerate to the fine-tuning set significantly reduces the failure rates on HateModerate and HateCheck, while the failure rates on the CardiffNLP's test sets are comparable. While adding **+HM** sometimes make the non-hate failure rate even worse than **CardiffNLP**, re-balancing the hateful and non-hateful examples can alleviate this problem. Furthermore, while OpenAI performs the best in Table 1 and Figure 4, in Table 2 it has higher failure rates than **+HM** and **+HM*** on the HateModerate test. This comparison with the strong OpenAI model further confirms the significance of our dataset.

Previous work has shown that fine-tuning hate speech models can lead to bias, e.g., the two-word sentence "*black women.*" is predicted as hateful (Markov et al., 2023; Reddit, 2023; Zhou et al., 2021; Davidson et al., 2019). We perform an analysis of measuring such bias in our fine-tuned model in Appendix A.10.

Finding Summary of RQ3. We find that by fine-tuning hate speech detection models with HateModerate, we can effectively reduce the models' non-conformity to content policies.

5 Conclusions

In this paper, we propose a dataset HateModerate, which includes hateful and non-hateful examples matching the 41 community standards guideline policies of Facebook. First, we leverage manual annotation with 28 graduate students followed by information retrieval, data augmentation, and verification to construct a dataset containing both hateful and non-hateful examples. Second, we use HateModerate to test the failures of state-of-the-art hate detection models. We find that popular content moderation models frequently make mistakes for both hateful and non-hateful examples. Finally, we observe that by augmenting the training data with HateModerate, the model can better conform to HateModerate while having the comparable performance to the original test data. Our study highlights the importance of investigating hate speech detectors' conformity to content policies.

6 Future Work

Extending Our Work to Any Natural Language Requirements. In this work, we focus on examining the models' performance against Facebook's policies. Although existing study shows that Facebook's content policies are more comprehensive, our dataset does not naturally generalize to other platforms' guidelines. One future direction is to enable the automatic retrieval of hateful and non-hateful examples matching any natural language requirements. The retriever needs to match a policy to specific examples by bridging the vocabulary gap while paying attention to subtle differences in the policy requirements, e.g., "*Dehumanizing as diseases→ XXX are cancer*".

Explaining Content Moderation Decisions. Linking a hate speech example to one of the policies can improve the accountability and transparency of automated hate speech detector. Our dataset can be used for the training and evaluation of this task.

7 Limitations

Extending HateModerate to New Policies. HateModerate is built based on Facebook's content moderation policy on Nov 23, 2022 (Facebook, 2022). When applying our work to different policies (e.g., for a different platform), we must hire new human annotators to search for the matching examples. One future direction for improving this limitation is to automatically retrieve the matching examples given the policy.

Comprehensiveness of Content Policies. Although Facebook's content moderation policies on hate speech are relatively comprehensive, the 41 policies may not cover all hate speech.

Mitigating the Data Bias of HateModerate. Our data collection leverages searches based on community standards guidelines. Since the searches are initiated based on the guidelines, the collected dataset may contain bias in the following aspects. First, the data might be skewed towards keywords explicitly mentioned or can be easily inferred from the guideline. Second, the dataset may contain limited *implicit hateful sentences*. One way to mitigate the first bias is to enumerate concepts given the high-level guideline, e.g., by querying the GPT engine: "*Enumerate a list of objects (i.e., things) for the dehumanization of women:* ". For the second bias, following previous work on implicit hateful examples (ElSherief et al., 2021), we plan to explore automated categorization to improve the coverage of implicit hate in HateModerate.

8 Ethics Considerations

License/Copyright. HateModerate primarily relies on reusing examples from existing hate speech data including DynaHate (Vidgen et al., 2021) and HateCheck (Röttger et al., 2021). We refer users to the original licenses accompanying each dataset.

Intended Use. HateModerate's intended use is as an evaluation tool for hate speech detection models, supporting capability tests to help diagnose model failures. We demonstrated this use of HateModerate in Section 4. We also briefly discussed alternative uses of HateModerate in Section 6, e.g., as a dataset for explaining a decision for hate moderation by linking the decision to one of the content policies. These uses aim at aiding the development of better hate speech detection models. HateModerate reuses existing hate speech datasets including DynaHate (Vidgen et al., 2021) and HateCheck (Röttger et al., 2021), and our usage for these datasets is consistent with the intended use described in their papers.

Potential Misuse. Similar to existing datasets for capability tests (Röttger et al., 2021), one potential misuse is over-extending claims about the functionalities of hate detection models. Our dataset may allow malicious actors to generative model that can generate hate speech matching the requirement for specific policies, which may further help them attack existing content moderators in a more structured manner. Nevertheless, due to the small scale of our dataset, this will unlikely happen. Overall, the scientific and social benefits of the research arguably outweigh the small risk of their misuse.

Annotator Recruitment and Compensation. HateModerate is annotated by 28 graduate students (10 Indian, 9 Chinese, 9 USA) in Computer Science, all of them are fluent English speakers. The student annotators in this paper are recruited from PhD and Master students at a research lab and students taking a graduate-level NLP course. They were rewarded $20 Amazon e-gift cards as compensation for their annotation efforts. The entire annotation process spans seven months while the actual annotation time takes about seven weeks (four weeks for hate, three weeks for non-hate). The annotator names are anonymized in the dataset. We obtained the annotators' consent, and it was explained to the annotators how the data would be used.

References

Surge AI. 2023. Twitter Hate Speech Data .

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The Workshop on Online Abuse and Harms (WOAH)*.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Workshop on Online Abuse and Harms*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*.

Cardiff NLP. 2023. CardiffNLP Twitter-RoBERTa-base-hate Model.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics*.

US Congress. 1996. Section 230 of the Communications Decency Act.

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*.

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 Conference on Web and Social Media*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Facebook. 2021. Facebook RoBERTa Model R1 for Hate Speech.

Facebook. 2022. Facebook Community Standards on Hate Speech .

Facebook. 2023. Community Standards Enforcement Report on Hate Speech Detection.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International Conference on Web and Social Media*.

Paul Friedl. 2023. Dis/similarities in the design and development of legal and algorithmic normative systems: the case of perspective api. *Law, Innovation and Technology*.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the Web Science Conference*.

Google. 2023a. A List of Publishers using the Perspective API .

Google. 2023b. Google Perspective API .

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 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*.

Ankita Gupta, Marzena Karpinska, Wenlong Zhao, Kalpesh Krishna, Jack Merullo, Luke Yeh, Mohit Iyyer, and Brendan O'Connor. 2023. ezCoref: Towards unifying annotation guidelines for coreference resolution. In *Findings of the Association for Computational Linguistics: EACL*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*.

X. He, S. Zannettou, Y. Shen, and Y. Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *Proceedings of the IEEE Symposium on Security and Privacy*.

Instagram. 2023. Instagram Community Guidelines .

Jialun 'Aaron' Jiang, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2020. Characterizing community guidelines on social media platforms. In *Conference Companion Publication of the Computer Supported Cooperative Work and Social Computing*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Conference on Artificial Intelligence*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the Conference on Artificial Intelligence*.

J Nathan Matias. 2016. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the CHI conference on human factors in computing systems*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

OpenAI. 2023. Embeddings - OpenAI API .

OpenAI. 2023a. OpenAI Moderation API .

OpenAI. 2023b. Using GPT-4 for content moderation .

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the International Workshop on Semantic Evaluation*.

Pinterest. 2023. Pinterest Community Guidelines .

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*.

Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. 2021. An information retrieval approach to building datasets for hate speech detection. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Reddit. 2020. Reddit Hate speech policy .

Reddit. 2023. Reddit Moderation API .

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the Workshop on Perspectivist Approaches to NLP @LREC2022*.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2020. Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. In *International Conference on Web and Social Media*.

Zeerak Talat. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the Workshop on NLP and Computational Social Science*.

Zeerak Talat 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*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca model .

Twitter. 2023. Twitter's Policies Hateful Conduct .

European Union. 2022. The Digital Services Act package.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Talat, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Workshop on Online Abuse and Harms*.

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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. TestAug: A framework for augmenting capability-based NLP tests. In *Proceedings of the International Conference on 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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Yuchen Zhou, Mark Dredze, David A. Broniatowski, and William D. Adler. 2019. Elites and foreign actors among the alt-right: The gab social media platform. *First Monday*.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop*.

A Appendix

A.1 Case Studies of Policies with High Failure Rates

- Tier 1: "*Dehumanizing speech as Certain Objectsects (women as household objects or property or objects in general; Black people as farm equipment; transgender or non-binary people as "it")*": In particular, most models fails to detect hate speeches such as "*MAKE WOMEN PROPERTY AGAIN!!!!*".

- Tier 2: "*Expressions of contempt, self-admission to intolerance on the basis of protected characteristics*": e.g., Perspective fail to detect "*I am racist because I'm a white person grown up in white American systems.*".

- Tier 3: "*Economic exclusion, which means denying access to economic entitlements and limiting participation in the labour market*": e.g., "*Pregnant women should not get any promotion at work as they will end up in maternity leave*".

- Tier 4: "*Content explicitly providing or offering to provide products or services that aim to change people's sexual orientation or gender identity*" We find all models fail in the half of this policy, e.g., "*With the right therapy and support gays can be cured and live much more fulfilling lives*".

A.2 The Classes of Non-Hateful Examples

- Counter hate or referencing hate, e.g., "*It is not correct to say 'Muslims should be killed'.*";

- Offensive towards non-protected entities, e.g., "*I hate cauliflower.*";

- Neutral, e.g., "*Black people have darker skins.*";

- Supportive, e.g., "*Women should be promoted.*";

- Mislabeled examples, i.e., hateful or the example does not match the policy;

A.3 The List of Evaluated Hate Speech Models

- Google's Perspective API: Jigsaw's Perspective API (Google, 2023b) is a popular model for content moderation, which is frequently used in downstream moderation tasks including news publishing, social media (Google, 2023a), as well as helping social and political science research (Friedl, 2023). Perspective leverages training data from a variety of sources, including comments from online forums such as Wikipedia and The New York Times[5].

- OpenAI's Moderation API: OpenAI's Moderation API (OpenAI, 2023a) OpenAI's content moderation endpoint, it is based on a GPT model fine-tuned using the classification head as the objective function (Markov et al., 2023). The fine-tuning leverages both public hate speech datasets and the production data of OpenAI, and it requires continuous training to adapt to the new hateful content (Markov et al., 2023). This model is being actively maintained and has been used by Stanford's Alpaca to improve the safety alignment of the text generation (Taori et al., 2023).

- Cardiff NLP's Fine-Tuned RoBERTa model: This open-source model is a fine-tuned RoBERTa model by Cardiff University's NLP group (Antypas and Camacho-Collados, 2023). The complete list of the 13 datasets used for fine-tuning can be found on the model's HuggingFace page: (Cardiff NLP, 2023). The older version of this model is the top-2 most downloaded fine-tuned model (84.6k downloads as of Oct 2023) for English hate-speech detection on the HuggingFace platform [6].

- Facebook's Fine-Tuned RoBERTa model (Facebook, 2021): This open-source model is a fine-tuned RoBERTa model by Facebook and the Alan Turing Institute (Facebook, 2021). The fine-tuning leverages 11 datasets, although the exact list is not revealed by the authors (Vidgen et al., 2021). The R4 version of this model is the top-1 most downloaded fine-tuned model (54k downloads as of Oct 2023) for English hate-speech classification on HuggingFace. Instead of R4, we evaluate the R1 model, because the R4 model is fine-tuned on DynaHate thus evaluating R4 causes the data contamination problem (Magar and Schwartz, 2022).

A.4 The List of the 9 Training Datasets for CardiffNLP's Model

Although the CardiffNLP model uses 13 datasets for fine-tuning (Antypas and Camacho-Collados, 2023), 4 datasets are non-downloadable, we list the 9 accessible datasets below:

- **Measuring hate speech (MHS)** (Sachdeva et al., 2022) include 39,565 social media comments.

- **Call me sexist, but (CMS)** (Samory et al., 2020) consist of 6,325 sentences related with sexism.

- **Hate Towards the Political Opponent (HTPO)** (Grimminger and Klinger, 2021) collect 3,00 tweets about the 2020 USA president election.

- **HateXplain** (Mathew et al., 2021) contains 20,148 posts from Twitter/X and Gab.

- **Offense** (Zampieri et al., 2019) is a collection of 14,100 tweets about offensive or non-offensive.

- **Automated Hate Speech Detection (AHSD)** (Davidson et al., 2017) combine 24,783 tweets.

- **Multilingual and Multi-Aspect Hate Speech Analysis (MMHS)** (Ousidhoum et al., 2019) is a dataset with 5,647 tweets in three different languages: English, Arabic and French.

- **HatE** (Basile et al., 2019) is a collection of 19,600 tweets with English and Spanish languages.

- **Detecting East Asian Prejudice on Social Media (DEAP)** (Vidgen et al., 2020) has 20,000 tweets which focus on East Asian prejudice.

A.5 Excluding Sentences to Prevent Data Contamination

In this paper, to reduce the risk of data contamination, i.e., overlaps between the train and test

[5]https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US

[6]https://huggingface.co/models?sort=downloads&search=hate

dataset, we need to exclude the examples from HateModerate that can potentially exist in the training data of the evaluated models. First, OpenAI API and Google Perspective have not released their training sets. Second, among the training datasets of CardiffNLP (Antypas and Camacho-Collados, 2023), we identify that Waseem et al. (Talat, 2016) and Founta et al. (Founta et al., 2018) are used in DynaHate's R0 dataset (Vidgen et al., 2021). As a result, we exclude all examples in DynaHate that are originally from other datasets and only keep those that are newly created. More specifically, we keep only the perturbed examples in rounds 2, 3, and 4. Finally, since Facebook's training datasets have no overlaps with the DynaHate, there is little risk of data contamination with HateModerate.

A.6 The Hypeparameters and Details of the Fine-Tuning Process

To study the effectiveness of HateModerate in reducing models' non-conformity issues, we fine-tune two RoBERTa models: ① Fine-tuning using the CardiffNLP 9 datasets in Section A.4; ② Fine-tuning using CardiffNLP datasets + HateModerate. The hyperparameter tuning process explores a range of learning rates and epoch sizes. Specifically, we experiment with grid search using the learning rates $1E-5, 2E-5$, epoch sizes $2, 3, 4$, and training batch size $4, 16, 32$. For both models, the warm-up steps are 50. The grid search space is chosen by referring to the best-performed hyperparameters setting of Cardiff NLP models as described in (Antypas and Camacho-Collados, 2023). The best-identified hyperparameters for both models are learning rate $= 2E-5$, batch size $= 32$, and epoch size $= 4$. Both models are fine-tuned on a server with 4x NVIDIA V100 GPUs, the training takes approximately half an hour per epoch for both models.

A.7 The Agreement Rates between Annotators

Table 3 includes detailed agreement rates on verifying whether an example matches or does not match a guideline. We report Krippendorf's α which is often used in previous work on crowdsourcing (Mathew et al., 2021; Vidgen et al., 2021) and the ratio of agreement.

Table 3: The inter-annotators agreement rates and Krippendorff's α in the HateModerate validation process.

HateModerate	Hate	Non-Hate
Ratio of Agreement	89.64%	91.15%
Krippendorff's α (Nominal)	0.521	0.808
Krippendorff's α (Interval)	0.521	0.809

A.8 Comparison of Model Failures of Different Sub-Categories of Non-Hateful Speeches

To better understand the failures in non-hateful examples, we further conduct a comparative study on the failure rates between different sub-categories of the non-hateful examples. We show the results in Figure 5. Among all the 4 non-hateful categories, we find that counter hate and attacking non-protected group has the highest failure rate, whereas advocating for protected groups has the lowest failure rate. This result is consistent with our expectation, since the former categories sound more aggressive.

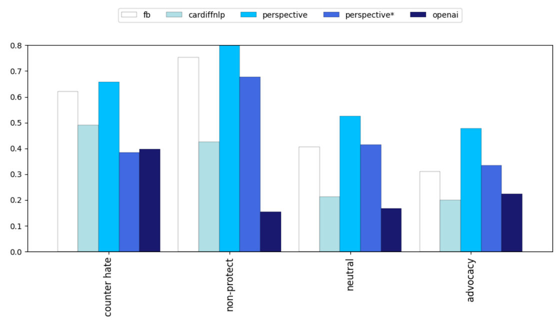

Figure 5: The comparison of failure rates in each subcategories of non-hateful examples

Table 4: Measuring the bias: Simple test cases in which we find the hate probability of the fine-tuned model abnormal. The numbers in this table are the hate probabilities for each example.

Test case	Cardiff	+HM	+HM*
Middle Eastern people are selfish.	**0.001**	**0.748**	**0.414**
Asians are selfish.	**0.795**	0.997	0.981
Homos are ugly.	0.986	**0.218**	**0.765**

Table 5: The statistics and p-values of the paired t-tests for comparing fine tuning with and without HateModerate

	Card vs **+HM**		**Card** vs **+HM***	
	statistics	p-value	statistics	p-value
***HateCheck* (*Röttger et al., 2021*)**				
Hate	**25.59**	**1.1E-133**	**20.43**	**3.0E-88**
Non-hate	-2.51	1.2E-02	-0.43	6.7E-1
Overall	**23.90**	**6.0E-118**	**18.09**	**2.7E-70**
HateModerate Test				
Hate	**20.79**	**2.9E-91**	**15.66**	**1.0E-53**
Non-hate	**5.85**	**5.4E-09**	**7.82**	**6.7E-15**
Overall	**12.11**	**3.7E-33**	**6.47**	**1.1E-10**
CardiffNLP Test Sets:				
***hatEval* (*Basile et al., 2019*)**				
Hate	1.18	2.4E-01	**3.31**	**9.4E-04**
Non-hate	-0.61	5.4E-01	-0.44	6.6E-01
Overall	1.19	2.4E-01	**2.17**	**3.0E-02**
***HTPO* (*Grimminger and Klinger, 2021*)**				
Hate	-2.32	2.1E-02	0.00	1.0E+00
Non-hate	0.73	4.7E-01	0.21	8.4E-01
Overall	-2.05	4.1E-02	-0.16	8.7E-01
***HateXplain* (*Mathew et al., 2021*)**				
Hate	-0.34	7.4E-01	-1.31	1.9E-01
Non-hate	**3.71**	**2.1E-04**	**3.63**	**2.9E-04**
Overall	-3.10	1.9E-03	-3.54	4.1E-04

A.9 Details on the Significance Tests for the Fine-Tuning Experiments

For the fine-tuning experiments in Table 2, we perform paired t-tests[7] between **CardiffNLP** vs **+HM** and **CardiffNLP** vs **+HM***. The statistics and p-values of the t-tests are shown in Table 5. For each t-test, if the statistics is positive, it means the CardiffNLP baseline performs better and vice versa. The results where **+HM** or **+HM*** significantly outperforms **CardiffNLP** are denoted in bold.

A.10 Measuring the Bias of the Fine-Tuned Models

Prior work shows that fine-tuning hate speech detectors can lead to bias against certain protected groups, e.g., the two-word sentence "*black women.*" is predicted as hateful. Previous work thus measures such bias when fine-tuning a hate detection model (Markov et al., 2023; Reddit, 2023). For example, Markov et al. (Markov et al., 2023) use 69k curated synthetic examples like "*[subject] is selfish/foolish/narrow-minded.*" to measure and mitigate the bias.

Following the previous work, we also measure the bias in this paper. We test the 3 fine-tuned models in Table 2 (i.e., **CardiffNLP**, **+HM**, **+HM***) with 21 synthetic non-hateful examples and 13 hateful examples such as "*[PROTECTED GROUP] ARE [POS/NEG ADJ].*". Surprisingly, almost 100% of the non-hateful predictions are correct while 15% of the hateful predictions are incorrect. In Table 4, we report the only test cases in which we find the hate probability of the fine-tuned model is abnormal.

Besides the simple examples in Table 4, we further measure the bias using more realistic examples from HateCheck (Röttger et al., 2021). HateCheck contains 18 hateful and 11 non-hateful suites of test cases on 7 protected groups. We find the 3 fine-tuned models generally have low failure rates on the non-hateful examples of HateCheck. In Table 6, we report all test suites in HateCheck whose failure rates are higher than 50%, including two test suites about women. To study whether adding HateModerate increases the bias compared to the original model, we further perform the paired t-test between **CardiffNLP** vs **+HM**'s predictions on HateCheck non-hateful examples (p-value: 0.80), and **CardiffNLP** vs **+HM*** (p-value: 0.83). Since the p-values are not significant, we can reject the null hypothesis that HateModerate introduces more bias to the model.

Table 6: Measuring the bias: all test suites in HateCheck whose failure rates are higher than 50%

Test Suite	Group	**Card**	**+HM**	**+HM***
F8: Non-hateful homonyms of slurs	Women	.80	.80	.70
F9: Reclaimed slurs	Women	.47	.67	.60
F23: Abuse targeted at individuals (not as member of a prot. group)	None	.45	.46	.52
F24: Abuse targeted at nonprotected groups (e.g. professions)	Non-Protected Group	.58	.52	.58

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

A.11 Overview of Facebook's Hate Speech Community Standards

In Table 7, we provide a copy of Facebook's community standards guidelines as of Nov 2022. The guidelines also serve as the text instructions provided to the human annotators for all six steps of data labeling in Section 3 (Figure 2).

Table 7: Short name and description for Facebook's Hate Speech Community Standards (Facebook, 2022). We show matching short names of guidelines and their index in Figure 4, the full descriptions of them are following.

ID	Tier	Guideline	Description
0	1	Dehumanize Filth	Dehumanizing speech: Filth (including but not limited to: dirt, grime)
1	1	Violent Speech	Violent speech or support in written or visual form
2	1	Dehumanize Insects	Dehumanizing speech: Insects (including but not limited to: cockroaches, locusts)
3	1	Dehumanize Bacteria	Dehumanizing speech: Bacteria, viruses, or microbes
4	1	Disease	Dehumanizing speech: Disease (including but not limited to: cancer, sexually transmitted diseases)
5	1	Dehumanize Animals	Dehumanizing speech: Animals in general or specific types of animals that are culturally perceived as intellectually or physically inferior (including but not limited to: Black people and apes or ape-like
6	1	Feces	Dehumanizing speech: Feces (including but not limited to: shit, crap)
7	1	Sexual Predator	Dehumanizing speech: Sexual predators (including but not limited to: Muslim people having sex with goats or pigs)
8	1	Subhumanity	Dehumanizing speech: Subhumanity (including but not limited to: savages, devils, monsters, primitives)
9	1	Criminal	Violent criminals (including but not limited to: terrorists, murderers, members of hate or criminal organizations). Other criminals (including but not limited to "thieves," "bank robbers," or saying "All [protected characteristic or quasi-protected characteristic] are 'criminals'").
10	1	Certain Objects	Certain objects (women as household objects or property or objects in general; Black people as farm equipment; transgender or non-binary people as "it")
11	1	Deny Existence	Statements denying existence (including but not limited to: "[protected characteristic(s) or quasi-protected characteristic] do not exist", "no such thing as [protected charactic(s) or quasi-protected characteristic]"), deny existence is different from contempt-should-not-exist in tier 2
12	1	Harmful Stereotype	Harmful stereotypes historically linked to intimidation, exclusion, or violence on the basis of a protected characteristic, such as Blackface; Holocaust denial; claims that Jewish people control financial, political, or media institutions; and references to Dalits as menial laborers
13	1	Hatecrime Victim	Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image.
14	2	Attack Hygiene	Content attacking a person or a group based on their hygiene, including but not limited to: filthy, dirty, smelly.
15	2	Attack Appearance	Content attacking a person or a group based on their physical appearance, including but not limited to: ugly, hideous.

16	2	Attack Intellectual Capability	Content attacking a person or a group based on their intellectual capacity, including but not limited to: dumb, stupid, idiots.
17	2	Attack Education	Content attacking a person or a group based on their education, including but not limited to: illiterate, and uneducated.
18	2	Attack Mental Health	Content attacking a person or a group based on their mental health, including but not limited to: mentally ill, retarded, crazy, insane.
19	2	Attack Characteristics	Attacking a person or group based on a perceived character trait, Content attacking a person or a group based on their character traits culturally perceived as negative, including but not limited to: coward, liar, arrogant, ignorant.
20	2	Attack Derogatory	Content attacking a person or a group based on derogatory terms related to sexual activity, including but not limited to: whore, slut, and perverts.
21	2	Less Than Adequate	Expressions about being less than adequate, including but not limited to: worthless, useless
22	2	Deviating Norm	Expressions about deviating from the norm, including but not limited to: freaks, abnormal.
23	2	Better Worse Than	Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
24	2	Contempt Self Admission Intolerance	Self-admission to intolerance on the basis of a protected characteristic, including but not limited to: homophobic, islamophobic, and racist.
25	2	Contempt Shouldn't Exist	Expressions that a protected characteristic shouldn't exist. (shouldn't exist is different from deny-existence in tier 1)
26	2	Contempt Despise Hate	Expressions of hate, including but not limited to: despise, hate.
27	2	Contempt Despise Dislike	Expressions of dismissal, including but not limited to: don´t respect, don't like, don´t care for
28	2	Curse Sexual	Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, Feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit.
29	2	Disgust Vomit	Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up.
30	2	Disgust Repulsive	Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck.
31	2	Curse Genitalia	Curse that referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole.
32	2	Curse Profane	Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker.
33	3	Segregation	Segregation in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting segregation.
34	3	Explicit Exclusion	Call for action of exclusion, e.g., explicit exclusion, which means things like expelling certain groups or saying they are not allowed.
35	3	Political Exclusion	Call for action of exclusion, e.g., political exclusion, which means denying the right to political participation.

36	3	Economic Exclusion	Call for action of exclusion, e.g., economic exclusion, which means denying access to economic entitlements and limiting participation in the labour market.
37	3	Social Exclusion	Call for action of exclusion, e.g., social exclusion, which means things like denying access to spaces (physical and online)and social services, except for gender-based exclusion in health and positive support Groups.
38	4	Change Sexual	Content explicitly providing or offering to provide products or services that aim to change people's sexual orientation or gender identity.
39	4	Attack Concepts	Content attacking concepts, institutions, ideas, practices, or beliefs associated with protected characteristics, which are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic.
40	4	Spread Virus	Content targeting a person or group of people on the basis of their protected characteristic(s) with claims that they have or spread the novel coronavirus, are responsible for the existence of the novel coronavirus, are deliberately spreading the novel coronavirus, or mocking them for having or experiencing the novel coronavirus.