HateModerate: Grounding and Benchmarking Hate Speech Detection with Content Policies

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

Anonymous EACL submission

Abstract

Social media platforms greatly facilitate user communications, but they also open the doors to unwanted contents such as hateful speech, misinformation, and pornography. To protect users from a massive scale of hateful contents, existing work investigate machine learning solutions for training automated hate speech moderators. Nevertheless, we identify that one gap is that few existing hate speech datasets are associated with a list of moderation rules. Without clarifying the moderation criteria, the trained moderator may behave differently from user's expectation. This work seeks to bridge this gap by creating a hate speech dataset matching a list of moderation rules. Using crowdsourcing, we search and collect a dataset named HateModerate grounded by Facebook's community standards guidelines for hate speech. We evaluate the performance of state-of-the-art hate speech detectors against HateModerate, revealing substantial discrepancies these models have with content policies. By fine-tuning one model with HateModerate, we observe that fine-tuning can effectively improve the models' conformity to policies. Our results highlight the necessity of developing rule-based datasets for hate speech detection. Our datasets and code can be found on: `https://sites.google.com/view/content-moderation-project`.

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, pornography, etc. The removal of such unwanted contents used to be handled by human moderators. In the recent years, thanks to the development of AI techniques, social media companies are actively investigating automated hate speech moderators powered by AI (fac, 2023; gpt); meanwhile, the ML/NLP research community are also vigorously developing new resources and improving the machine learning techniques for automated hate speech detection (Waseem and Hovy, 2016; Waseem, 2016; Davidson et al., 2017; Founta et al., 2018; Vidgen et al., 2020b; Röttger et al., 2020; Mathew et al., 2021; He et al., 2021; ElSherief et al., 2021; Hartvigsen et al., 2022; Sachdeva et al., 2022; Markov et al., 2022; Antypas and Camacho-Collados, 2023). Following the works, researchers published language models fine-tuned with these resources to facilitate downstream moderation tasks (per, b; ope; car; fb, c).

Hate Speech Community Standards Guidelines

Tier 1:
Content targeting a person or group of people on the basis of their protected characteristic(s) with:

Dehumanizing Speech
Compare the protected groups as animals that are perceived as inferior (*including but not limited to: apes, pigs*)
Compare the protected groups as feces (*including but not limited to: shit, crap*)
… …
Violent Speech
Advocacy for physical harm to protected groups (*including but not limited to: beat up, kill*)
Threats of weaponry to protected groups (*including but not limited to: shoot, stab*)
… …

Figure 1: An example of community standards guidelines for hate speech (fb, a)

Nevertheless, there exist one aspect that, to the best of our knowledge, was neglected by existing work in hate speech detection. That is, the existing datasets are not grounded by a list of rules or criteria for what speeches are considered as hateful. The criteria of hate speech often vary according to the moderation needs. For example, Gab allows more elitism speeches than Twitter (gab). Similarly, the labels in the existing hate speech datasets may or may not conform to the same criteria as where the trained detector is being deployed to. Without clarifying the rules, the hate speech detector may behave differently from expectation, which undermines its accountability. The closest work to a rule-based dataset is HateCheck (Röttger et al., 2020), but their rules focus on the syntactic structures, thus they suffer from a low coverage on the hate speech categories (Section 4.3 of (Röttger et al., 2020)).

To improve the accountability of automated content moderators, this paper proposes a dataset called HateModerate, which consists of a list of test suites containing hateful and non-hateful examples matching content moderation rules. Among the published moderation rules from existing work (Banko et al., 2020; fb, a; Röttger et al., 2020), we opt for Facebook's community standards guidelines for hate speech (fb, a) as previous work shows it is the most comprehensive among all platforms (Jiang et al., 2020) and it has good clarity. Two examples of Facebook's guidelines are shown in Figure 1.

HateModerate is collected using the process below. First, crowdsourced annotators are instructed to manually search for hateful examples from existing datasets matching each policy. The process is followed by a validation step to ensure the label accuracy. After the hateful examples are collected, we retrieve difficult non-hateful examples that closely resemble the hateful examples in each policy which helps improve the detection of model failures. We further validate the non-hateful examples by leveraging a human-LLM collaborative annotation process. The average agreement rate for the hateful examples is 87% and for non-hateful examples is 88%.

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 (per, b), OpenAI's Moderation API (ope), Facebook's RoBERTa model (fb, b) and Cardiff NLP's RoBERTa model (car). 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, especially for counter hate and attacking non-protected entities.

After observing the model failures, we further seek answers to how to improve model conformity to policies. To this end, we compare the results of two models: first, we fine-tune a RoBERTa model using the training datasets of the CardiffNLP model; second, we fine-tune a RoBERTa model using CardiffNLP's training data and HateModerate. We find that compared to the first model, the second model consistently reduces the model's failures on HateModerate, while maintaining the same performance on the original testing data of CardiffNLP. This result shows that including a rule-based training set can effectively alleviate the model's nonconformity issue to policies, which underscores the importance of keeping the dataset grounded with the moderation criteria.

2 Background and Related Work

In this section, we introduce the background on hateful content moderation and NLP model evaluation, which helps explain the motivation of our work.

2.1 Automated Content Moderation

The removal of hateful contents online is an important process for keeping social media platforms safe and healthy, as well as reducing the incitement of real-world harms (un). Due to the difficulty of understanding hateful contents, social media platforms largely relied on human moderators for content removal. Recently, companies such as Facebook and OpenAI have investigated automated content moderation powered by NLP techniques to scale up the moderation process and to alleviate human moderators' workload (fac; Markov et al., 2022). For example, Facebook deployed a fine-tuned multilingual RoBERTa model and a hybrid system to moderate the hate speech on Facebook (fac, 2023; eve). OpenAI also fine-tuned a GPT model with classification loss for moderating harmful contents in their products (Markov et al., 2022). They found the model must be continuously updated to adapt to the new hateful contents (Markov et al., 2022).

Improving Machine Learning for Hate Speech Detection. Alongside the companies' efforts, the hate speech community has released multiple public labeled hate speech datasets for training machine learning models (Waseem, 2016; Waseem and Hovy, 2016; Davidson et al., 2017; Golbeck et al., 2017; Founta et al., 2018; Hartvigsen et al., 2022; Vidgen et al., 2020b). These datasets allow researchers to fine-tune models to a diverse range of hateful examples and thus can potentially generalize better to unseen examples. For example, OpenAI combined public datasets and their production data to train the initial model of their Moderation API endpoint before continual learning (ope; Markov et al., 2022). Both Cardiff University's NLP lab and Facebook fine-tuned an open-source RoBERTa model to a list of selected public datasets (Facebook used 11 while CardiffNLP used 13), which rank top-2 and top-1 among the most downloaded hate detection models on HuggingFace (Vidgen et al., 2020b; Antypas and Camacho-Collados, 2023). To this day, fine-tuning remains the state-of-the-art technique for training automated hate detectors, and the fine-tuned models are used in real-world downstream moderation tasks (alp).

2.2 Policies and Rules for Content Moderation

Issues with Existing Models. One issue with fine-tuning public datasets for hate speech (Vidgen et al., 2020b; Antypas and Camacho-Collados, 2023) is that their moderation criteria is not entirely clear. Essentially, what speeches are considered hateful vary across platforms. For example, Gab allows more elitism speeches than Twitter (gab). When fine-tuning public datasets, it is thus unclear whether these datasets labels are consistent with the user's own application scenario.

Grounding Hate Speech Datasets with Rules/Labels. To explain the criteria of hatefulness, existing work has associated fine-grained labels with each hateful example in the dataset. For example, DynaHate (Vidgen et al., 2020b) and Measuring Hate Speech (Sachdeva et al., 2022) label each example with fine-grained categories such as derogatory, dehumanization, and insult. However, these categories are high-level concepts and it is difficult to follow them as the labeling rules, e.g., it is difficult to search hateful examples matching the rule "*insult*".

Taxonomies/Rules/Policies for Content Moderation. Another line of existing work construct taxonomies for content moderation (Banko et al., 2020; fb, a; Röttger et al., 2020). A taxonomy contains a list of rules, each specified by a natural language description. For example, Banko et al. (Banko et al., 2020) introduces a taxonomy for various unwanted contents, e.g., sexual aggression, doxxing, misinformation. HateCheck (Röttger et al., 2020) provides a list of rules for hate speech. Nevertheless, most of the rules of HateCheck focus on defining hate speeches with syntactic structures rather than semantic meanings, and HateCheck's rules suffer from a low coverage on the hate speech categories, which is explained in Section 4.3 of (Röttger et al., 2020).

Community Standards Guidelines. Community standards guidelines are policies on what contents are prohibited on social media platforms. Recently, major platforms all released their own guidelines, e.g., Twitter (twi, b), Instagram (ig), and YouTube (yt). Jiang et al. (Jiang et al., 2020) conducted a comparative study for the existing community standards guidelines across platforms, their study suggests that Facebook's guidelines are the most comprehensive ones above all.

Facebook provides a list of 41 community standards guidelines for hate speech moderation (fb, a). Since each guideline is a natural language specification of hate speech, the guidelines can be used as a taxonomy for defining the moderation criteria of the dataset. Figure 1 shows two of Facebook's hate speech guidelines and Table 3 shows the complete list. These guidelines are organized into 4 tiers based on content severity (fb, a): Tier 1 includes the most offensive content, e.g., dehumanization and violence towards protected groups; Tier 2, Tier 3, and Tier 4 are less severe, e.g., stereotyping and contempts towards protected groups. From Figure 1 and Table 3 we can observe that Facebook's guidelines include *detailed specifications by enumerating specific examples of verbs and nouns*. Compared to other taxonomies, the detailed descriptions make it easy to identify the matched examples using keywords search. In this work, we thus leverage Facebook's community standards guidelines for constructing a dataset grounded by moderation rules.

2.3 Benchmarking NLP Model Performance with Capability Tests

Traditionally, NLP models are evaluated using the held-out mechanism, i.e., using data from the same distribution for training and testing. However, the in-distribution evaluation may overestimate the performance of a biased model (Belinkov et al., 2019). To examine whether the model has actually achieved the desired capabilities for the task, existing work constructs *capability tests* (Ribeiro et al., 2020; Röttger et al., 2020; Yang et al., 2022), i.e., out-of-domain test suites for benchmarking the models' capabilities under the task. In particular, HateCheck benchmarked the performance of 3 hate detection models (Google Perspective, Two Hat's SiftNinja and BERT) using 29 test suites for hate and non-hate capabilities. In this work, we propose HateModerate to benchmark models' capabilities in understanding hate speech conforming to hate policies.

3 Constructing the HateModerate Dataset

To bridge the gap in existing work on grounding hate speech detection datasets with moderation criteria, we propose a dataset, HateModerate, which consists of a list of test suites, each contains hateful and non-hateful examples matching one of Facebook's community standards guidelines of hate speech (fb, a) (Table 3). In this section, we describe the steps for the construction of HateModerate.

Human Annotators. HateModerate is annotated by 9 graduate students (4 Indian, 3 Chinese, 2 USA) in Computer Science, all of them are fluent English speakers and have taken at least one NLP course before. The annotation process is overseen by two experts in online hate. The annotation process take approximately 7 weeks. All participants are compensated with gift cards. The annotator names are anonymized in the dataset. We obtained annotators' consent and it was explained to the annotators how the data will be used.

Data Sources. In this work, instead of collecting new examples, we reuse existing examples from public datasets. This is because existing public datasets already provide good coverage of the common discourse of hate speech; reusing previously acclaimed public databases significantly reduces the workload and minimizes newly introduced annotation errors. In particular, we leverage the following 8 datasets: DynaHate (Vidgen et al., 2020b), Toxic Spans (Pavlopoulos et al., 2021), Hate Offensive (Davidson et al., 2017), HateCheck (Röttger et al., 2020), Twitter Hate Speech (twi, a), Ethos (Mollas et al., 2020), FRENK (Ljubešić et al., 2019), and COVID Hate and Counter Speech (Ziems et al., 2020). The hate/non-hate labels are available in all datasets.

3.1 Collecting Hateful Examples

Initial Manual Matching. 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 from the 8 datasets above. If insufficient, they can manually create or use chatGPT/GPT-3 to generate synthetic examples. Synonyms and regular expressions are employed to enhance the search efficiency. For example, for Guideline 0 (Tier 1, dehum filth), the annotator uses the regular expression ".*(filth|dirt).*". Multiple annotators also report that they seek for help from Google, ChatGPT and other team members to correctly understand the policy. For example, for Guideline 13 (Tier 1, hatecrm vic), the assigned annotator first struggles to find enough examples because the exact word "*hate crime*" rarely appear in any example, after a discussion, the team members suggest him to search for concrete hate crime keywords including *lynching* and *holocaust*. He is able to add more examples as a result.

Problems with the Initial Manual Matching. After the initial matching, we find a significant amount of falsely matched examples. The main reasons are of two folds. First, annotators interpret the policy criteria differently. For example, for Guideline 28 (Tier 2, curs sexual), the examples initially identified by the annotator only contain the curse words themselves but do not call for sexual activities. Second, when two policies look similar, it is easy to confuse between them, e.g., Guideline 11 (Tier 1, deny exist) vs Guideline 25 (Tier 2, cont shldnt exist). The team discuss and clarify the meanings of these policies.

Validating Hateful Examples. Due to the problems with the initial matching, we include a second stage of annotation. For each policy, we ask two additional annotators other than the initial annotator to assess whether the initial sentence match the policy descriptions, labeling them as 1 (valid) or 0 (invalid). Following existing work on using demonstrated examples to improve the quality of crowdsourced annotations (Gupta et al., 2022), we provide a few falsely-match examples for each policy, e.g., for Guideline 29 (Tier 2, curs sexual), "*They make me so pissed off these immigs!!*" is demonstrated as a false match. Between the two annotators for the validity, the average agreement rate over all 41 policies is 87%.

To minimize the mismatch with policies in HateModerate, we remove all examples with at least one invalid label. After validation, 4,600 examples remain, and some policies contain too few examples. To augment these policies, one additional annotator is asked to add more examples until no other examples can be found from the 8 datasets.

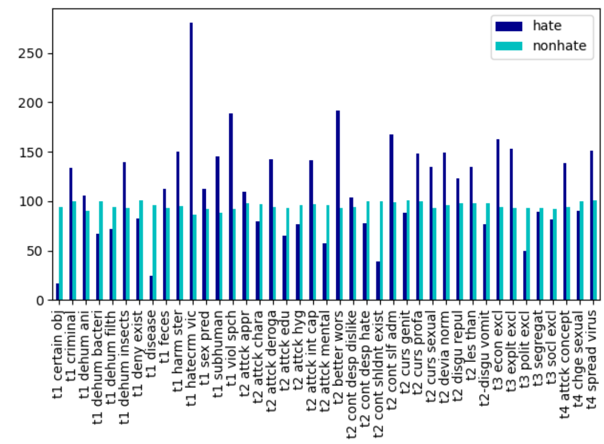

Figure 2: The statistics of examples in each policy in our dataset

3.2 Collecting Non-Hateful Examples

Retrieving Difficult Non-Hateful Examples. Testing with only hateful example will result in bias (e.g., one model has low failure rate simply because it sets a low threshold for hate), we further add non-hateful examples to HateModerate. 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. To this end, the corpus we retrieve from are all the non-hateful examples in DynaHate (Vidgen et al., 2020b), as a large proportion of DynaHate are manually perturbed examples. The retrieval algorithm follow the state-of-the-art dense retrieval paradigm (Karpukhin et al., 2020). We employ OpenAI's Embedding API (Ope) 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 HateModerate.

Classification of Non-Hateful Examples. After retrieval, we identify some mismatched non-hateful examples and mislabeled hateful examples. To remove them, 6 annotators further manually label each non-hateful examples into one of 5 fine-grained classes including counter hate, neutral, and mismatched examples. The full descriptions of the 5 classes can be find in Appendix A.2.

Validating Non-hateful Cases. After the initial manual classification, we find that some annotators confuse between the 5 classes. Inspired by previous work that leverages human-GPT collaboration to improve crowd-sourced labeling (He et al., 2023), we employ GPT-4 to generate a reference class from 1-5[1]. Subsequently, the original human annotator is asked to revisit all inconsistent cases and update their initial labels if they alter their opinion. After this validation stage, there remain 11.78% disagreement between human and GPT-4. For these inconsistent cases, the expert annotators re-evaluate and re-label them by referring to both GPT-4 and the original annotators' labels.

3.3 Dataset Statistics

In our final HateModerate dataset, we compile 6,826 examples (4,651 hateful, 2,175 non-hateful). It's important to note that some instances are duplicated because a single sentence can fall under multiple guidelines simultaneously. The majority examples come from DynaHate (Vidgen et al., 2020b) (4,979), followed by HateCheck (442), Toxic Span (100), GPT (762), manual (257), COVID hate (152), Hate Offensive (91), Ethos (11), Twitter Hate (33), and FRENK (19).

Figure 2 shows the statistics of HateModerate by policy. Among the 41 policies, the most frequent policy contains 361 examples whereas the least frequent policy contains 111 examples, most policies contain 100 to 250 examples, and the majority policies contain more than 100 examples.

4 Benchmarking Hate Speech Detectors' Consistency with Content Policies

In this section, we employ HateModerate as our evaluation benchmark to assess how AI-based hate speech 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 fine-tuning HateModerate improve models' conformity to content policies?

4.1 Experiment Setup

Hate Speech Models Evaluated. To answer RQ1-RQ3, we evaluate state-of-the-art models from

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

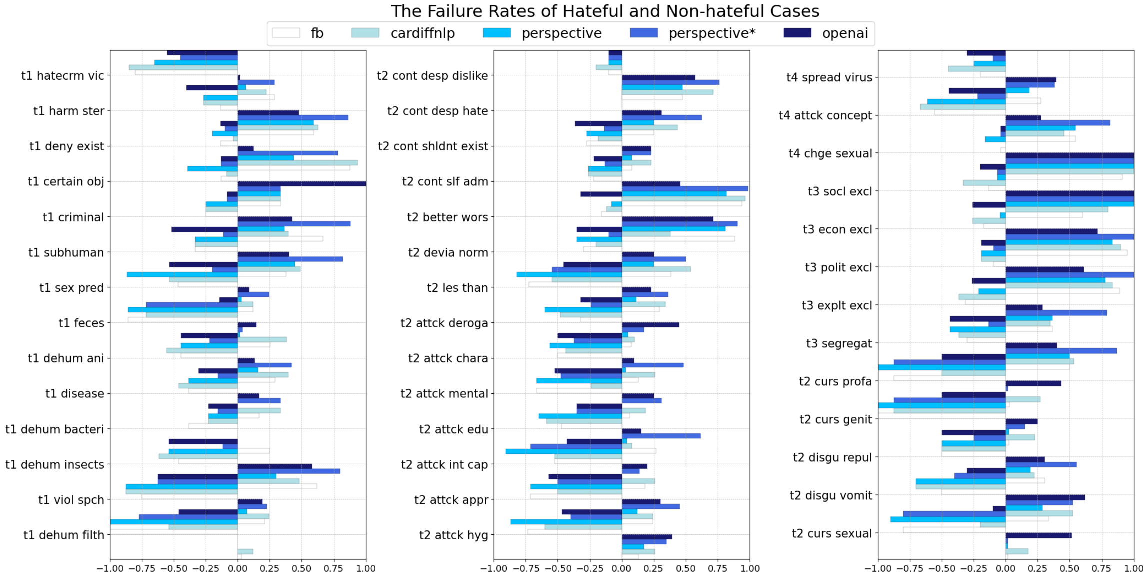

Figure 3: We detect the failure rates for both hateful and non-hateful examples across each of the 41 policies in Facebook's community standards guidelines (fb, a). Perspective's threshold is 0.5; Perspective*'s threshold is 0.7. For each policy, the bars facing right show the failure rates of hateful 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	.36	.36	**.20**	.43	.27	.43	.47	.45	.52	**.27**	.36	.67	.61	.62	.69	**.75**	.43	.44	.42	.52	**.34**
2	.33	.27	.34	**.20**	.43	.35	.48	.49	.40	.58	.38	**.36**	.65	.68	.63	**.70**	.57	.44	.47	.39	.55	**.35**
3	.65	.66	.68	.70	.93	**.60**	.24	.20	.30	.19	**.06**	.27	.38	.31	.32	**.45**	.43	.29	.26	.30	.37	**.22**
4	.55	.58	**.49**	.58	.73	.56	.33	.27	.37	.34	**.12**	.26	.49	.48	**.61**	.48	.38	.29	.24	.32	.39	**.20**

both industry API endpoints and open-source hate speech detection models. For industry APIs, we choose Google's Perspective API (per, b) and OpenAI's Moderation API (ope; Markov et al., 2022), which are frequently used in downstream detection tasks (alp; per, a); for open-source models, we choose Cardiff NLP's fine-tuned RoBERTa model (car) and Facebook's Fine-Tuned RoBERTa model (fb, b) which rank top-2 and top-1 among the most downloaded hate models on HuggingFace (hug). The full details of the models can be found in Appendix A.3.

Further Processing. To answer RQ3, we reserve half 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 comparison 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[2]. 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., 2020; 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 3. In addition, we report the average failure rate and the average toxicity score of each model in each tier in Table 1. From Figure 3 and Table 1 we can make the following observations.

[2]As a result, some guidelines do not have enough cases in Figure 3 so we skip them.

4.2.1 Comparison of Model Failures of different Policies

From Figure 3 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: 0.33, std: 0.13), followed by Perspective (avg: 0.34, std: 0.15). Facebook's RoBERTa (avg: 0.37, std: 0.15) and CardiffNLP (avg: 0.38, std: 0.16) 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 tends to assign higher toxicity for both hateful and non-hateful examples. As a result, the thresholds for Perspective should be higher than 0.5.

4.2.3 Comparison of Model Failures of Different Sub-Categories of Non-Hateful Speeches

In this section, 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 4. 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.

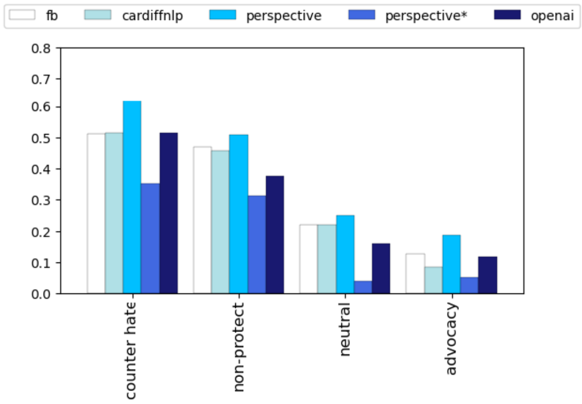

Figure 4: The comparison of failure rates in each sub-categories of non-hateful examples

Finding Summary of RQ1 and RQ2. ① For higher tier policies, the failure rates for hateful examples are lower and for non-hateful examples are higher; ② Among all models, 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. Among all difficult non-hateful examples, counter-hate is the most difficult whereas supporting protected groups is the easiest.

4.3 Mitigating Model Failures with Fine-Tuning HateModerate

In this section, we seek the answer to RQ3. We do so by comparing the results of the two models: ① A RoBERTa-base model fine-tuned using all the available training data for the CardiffNLP model (Antypas and Camacho-Collados, 2023)[3]; ② A RoBERTa-base model fine-tuned using

[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.

Table 2: Fine-tuning the RoBERTa Base Model on CardiffNLP training datasets with and without HateModerate.

Test / FailureRate	RoBERTa Fine-tuned on	
	CardiffNLP	+ HateModerate
***HateCheck* (*Röttger et al., 2020*)**		
Hate	57.50%	**37.42%**
Non-hate	**15.70%**	16.51%
Overall	44.14%	**30.76%**
HateModerate Test		
Hate	49.13%	**23.44%**
Non-hate	**15.39%**	22.03%
Overall	41.40%	**23.21%**
CardiffNLP Test Sets:		
***hatEval* (*Basile et al., 2019*)**		
Hate	**9.05%**	9.29%
Non-hate	79.31%	**78.79%**
Overall	49.80%	**49.60%**
***HTPO* (*Grimminger and Klinger, 2021*)**		
Hate	**71.19%**	76.27%
Non-hate	1.85%	**1.84%**
Overall	**8.67%**	9.17%
***HateXplain* (*Mathew et al., 2021*)**		
Hate	**17.25%**	17.60%
Non-hate	29.28%	**27.49%**
Overall	22.14%	**21.62%**

CardiffNLP's training data + HateModerate's reserved training data. We opt against continuously fine-tuning the original CardiffNLP model to HateModerate since the continuous fine-tuning is known to be prone to catastrophic forgetting (French, 1999). For the 9 training datasets of CardiffNLP model, we use the same train/test split as the original datasets[4]. The detail of the fine-tuning process can be found in Appendix A.6.

Results of Fine-Tuning. In Table 2, we compare the failure rates of the two fine-tuned models on the following test collections: ① The testing fold of HateModerate; ② The 3 testing datasets of CardiffNLP; ③ HateCheck (Röttger et al., 2020), a dataset for independent out-of-domain capability tests of hate speech. 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. The fine-tuning experiments show that adding HateModerate can effectively reduce hate detection models' conformity issue to content policies.

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.

[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 train.

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 (fb, a). We opt for study of Facebook guidelines due to its comprehensiveness (Jiang et al., 2020) and the high clarity of the guidelines. First, we leverage crowdsourcing followed by manual validation to construct a quality dataset for test cases of both hateful and non-hateful examples matching each policy. Second, we use HateModerate to test state-of-the-art hate detection models' conformity to the policies. We find that the most popular content moderation models (e.g. FB, CardiffNLP, OpenAI and Google) frequently make mistakes for both hateful and non-hateful examples. Finally, we observe that fine-tuning hate detection models with HateModerate can effectively reduce models' non-conformity issues to content policies. Our study underscores the importance of maintaining a set of rules for training and testing the performance of AI-based hate speech detectors.

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 than others (Jiang et al., 2020), our model 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 difference 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

Cost of Manual Annotation. HateModerate is built based on Facebook's content moderation policy on Nov 23, 2022 (fb, a). When applying our work on different policies (e.g., for a different platform), we must hire new human annotators. One of possible solution we tried in non-hateful part is the utilization of auto-labeling techniques by large language models.

Comprehensiveness of Policy Requirements. Although Facebook's content moderation policies on hate speech are relatively comprehensive, the 41 policies may not completely cover all hate speeches.

Contexts and User Expectation of Hate Speech. Our study focuses on checking AI-based content moderation software's behavior against policies. When evaluating the moderation software, we have not considered the context. However, whether a sentence is hateful or not may depends on the context; the same sentence may sounds hateful in one context but not in another. Moreover, the rules in content moderation policies may not exactly match user's expectation.

8 Ethics Considerations

License/Copyright. HateModerate primarily relies on reusing examples from existing hate speech data including DynaHate (Vidgen et al., 2020b) and HateCheck (Röttger et al., 2020). 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., 2020b) and HateCheck (Röttger et al., 2020), and our usage for these datasets is consistent with the intended use described in their papers.

Potential Misuse. Similar as existing datasets for capability tests (Röttger et al., 2020), one potential misuse is overextending 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 outweighs the small risk of their misuse.

Annotator Compensation. The student annotators in the project were rewarded giftcards compensations for their annotation efforts.

References

a. A List of Publishers using the Perspective API .

CardiffNLP Twitter-RoBERTa-base-hate Model.

Embeddings - OpenAI API .

Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content.

a. Facebook Community Standards on Hate Speech .

b. Facebook RoBERTa Model R1 for Hate Speech.

c. Facebook RoBERTa Model R4 for Hate Speech.

Gab displays less conversation, and more elitism, than Twitter .

b. Google Perspective API .

c. Google Perspective's Training Data .

Hate speech and real harm .

HuggingFace's most downloaded models for hate speech detection.

Instagram Community Guidelines .

OpenAI Moderation API .

Stanford Alpaca model .

a. Twitter Hate Speech Data .

b. Twitter's Policies Hateful Conduct .

Update on Our Progress on AI and Hate Speech Detection.

Using GPT-4 for content moderation .

YouTube Hate speech policy .

2023. Community Standards Enforcement Report on Hate Speech Detection.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. *arXiv preprint arXiv:2307.01680*.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*, pages 125–137.

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 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. *arXiv preprint arXiv:1907.04389*.

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.

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. *arXiv preprint arXiv:2109.05322*.

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 AAAI Conference on Web and Social Media*, volume 12.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

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*, 15(1):25–59.

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 2017 ACM on web science conference*, pages 229–233.

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.

Ankita Gupta, Marzena Karpinska, Wenlong Zhao, Kalpesh Krishna, Jack Merullo, Luke Yeh, Mohit Iyyer, and Brendan O'Connor. 2022. ezcoref: Towards unifying annotation guidelines for coreference resolution. *arXiv preprint arXiv:2210.07188*.

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. *arXiv preprint arXiv:2203.09509*.

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 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

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 2020 on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 287–291, New York, NY, USA. Association for Computing Machinery.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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*, pages 103–114. Springer.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2022. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274*.

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 AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

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

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

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 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

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*.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020a. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020b. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

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 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.

Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. Testaug: A framework for augmenting capability-based nlp tests. *arXiv preprint arXiv:2210.08097*.

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.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.

A Appendix

A.1 Case Studies of Policies with High Failure Rates

- Tier 1: "*Dehumanizing speech as certain objects (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 a 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, 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 (per, b) is a popular model for content moderation, which is frequently used in downstream moderation tasks including news publishing, social media (per, a), 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 (per, c).

- OpenAI's Moderation API: OpenAI's Moderation API (ope) 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., 2022). 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 contents (Markov et al., 2022). This model is being actively maintained and has been used by Stanford's Alpaca to improve the safety alignment of the text generation (alp).

- Cardiff NLP's Fine-Tuned RoBERTa model: This open-source model is a fine-tuned RoBERTa model by Cardiff University's NLP group (car). The complete list of the 13 datasets used for fine-tuning can be found on the model's HuggingFace page: (car). 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 (hug).

- Facebook's Fine-Tuned RoBERTa model (fb, b): This open-source model is a fine-tuned RoBERTa model by Facebook and the Alan Turing Institute (fb, b). The fine-tuning leverages 11 datasets, although the exact list is not revealed by the authors (Vidgen et al., 2020b). 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 (car), 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., 2020a) 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 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 (car), we identify that Waseem et al. (Waseem, 2016) and Founta et al. (Founta et al., 2018) are used in DynaHate's R0 dataset (Vidgen et al., 2020b). As a result, we exclude all examples in DynaHate which are originally from other datasets and only keep those that are newly created. More specifically, we keep only the perturbed examples in round 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 model: ① Fine-tuning using the CardiffNLP 9 datasets in Section A.4; ② Fine-tuning using CardiffNLP datasets + HateModerate. For both models, we use a training batch size of 32, a learning rate of $5E-6$, and an epoch size of 2. Both models are fine-tuned on a server with 4x V100 GPUs, the training takes approximately 1 hour for both models.

A.7 Overview of Facebook's Hate Speech Community Standards

Table 3: Short name and description for Facebook's Hate Speech Community Standards (fb, a). We show matching short names of guidelines and their index in Figure 3, the full descriptions of them are following.

ID	Tier	Guideline	Description
0	1	dehum filth	Dehumanizing speech: Filth (including but not limited to: dirt, grime)
1	1	viol spch	Violent speech or support in written or visual form
2	1	dehum insects	Dehumanizing speech: Insects (including but not limited to: cockroaches, locusts)
3	1	dehum bacteri	Dehumanizing speech: Bacteria, viruses, or microbes
4	1	disease	Dehumanizing speech: Disease (including but not limited to: cancer, sexually transmitted diseases)
5	1	dehum ani	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	sex pred	Dehumanizing speech: Sexual predators (including but not limited to: Muslim people having sex with goats or pigs)
8	1	subhuman	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 obj	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 exist	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	harm ster	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	hatecrm vic	Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image.
14	2	attck hyg	Content attacking a person or a group based on their hygiene, including but not limited to: filthy, dirty, smelly.
15	2	attck appr	Content attacking a person or a group based on their physical appearance, including but not limited to: ugly, hideous.
16	2	attck int cap	Content attacking a person or a group based on their intellectual capacity, including but not limited to: dumb, stupid, idiots.
17	2	attck edu	Content attacking a person or a group based on their education, including but not limited to: illiterate, uneducated.
18	2	attck mental	Content attacking a person or a group based on their mental health, including but not limited to: mentally ill, retarded, crazy, insane.

19	2	attck chara	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	attck deroga	Content attacking a person or a group based on derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts.
21	2	les than	Expressions about being less than adequate, including but not limited to: worthless, useless
22	2	devia norm	Expressions about deviating from the norm, including but not limited to: freaks, abnormal.
23	2	better wors	Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
24	2	cont slf adm	Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist.
25	2	cont shldnt exist	Expressions that a protected characteristic shouldn't exist. (shouldn't exist is different from deny-existence in tier 1)
26	2	cont desp hate	Expressions of hate, including but not limited to: despise, hate.
27	2	cont desp dislike	Expressions of dismissal, including but not limited to: don´t respect, don't like, don´t care for
28	2	curs 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	disgu vomit	Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up.
30	2	disgu repul	Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck.
31	2	curs genit	Curse that referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole.
32	2	curs profa	Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker.
33	3	segregat	Segregation in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting segregation.
34	3	explt excl	Call for action of exclusion, e.g., explicit exclusion, which means things like expelling certain groups or saying they are not allowed.
35	3	polit excl	Call for action of exclusion, e.g., political exclusion, which means denying the right to political participation.
36	3	econ excl	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	socl excl	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	chge sexual	Content explicitly providing or offering to provide products or services that aim to change people's sexual orientation or gender identity.
39	4	attck concept	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.