Detecting Community Sensitive Norm Violations in Online Conversations

Chan Young Park[♣] **Julia Mendelsohn**[♠] **Karthik Radhakrishnan**[♣] **Kinjal Jain**[♣]
Tushar Kanakagiri[♣] **David Jurgens**[♠] **Yulia Tsvetkov**[♡]
[♣]Language Technologies Institute, Carnegie Mellon University
[♠] University of Michigan
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{chanyoun,kradhak2,kinjalj,tkanakag}@cs.cmu.edu,
{juliame,jurgens@umich.edu}, yuliats@cs.washington.edu

Abstract

Online platforms and communities establish their own norms that govern what behavior is acceptable within the community. Substantial effort in NLP has focused on identifying unacceptable behaviors and, recently, on forecasting them before they occur. However, these efforts have largely focused on toxicity as the sole form of community norm violation. Such focus has overlooked the much larger set of rules that moderators enforce. Here, we introduce a new dataset focusing on a more complete spectrum of community norms and their violations in the local conversational and global community contexts. We introduce a series of models that use this data to develop context- and community-sensitive norm violation detection, showing that these changes give high performance.[1]

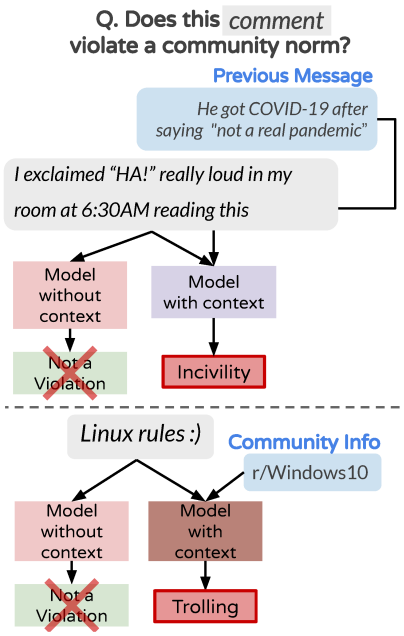

Figure 1: Two example comments[3] that were moderated due to violating community norms. The examples highlight the importance of contexts (i.e. conversation history and community information) in detecting community norm violation.

1 Introduction

Online communities establish their own norms of what is acceptable behavior (Danescu-Niculescu-Mizil et al., 2013; Jhaver et al., 2018; Rajadesingan et al., 2020). These norms run the gamut from *no hate speech* or *no personal attacks* to more idiosyncratic expectations of *content formatting* and *content sharing* (Chandrasekharan et al., 2018; Fiesler et al., 2018). Community moderators are responsible for identifying and removing rule-breaking content, regardless of whether users violate rules intentionally or unintentionally due to unfamiliarity with community norms.

Moderators of online communities often face a tough challenge of triaging the massive flow of content (Kiene et al., 2016; Dosono and Semaan, 2019; Kiene et al., 2019); for example, over 2 billion comments were posted to Reddit in just 2020.[2] Moderators have looked to technology to help support their role, using regex-based tools like Automoderator to flag potentially rule-breaking comments (Jhaver et al., 2019). Prior work has aimed to assist by developing machine learning techniques to recognize unacceptable content—yet these have focused on only the most socially-harmful violations, such as hate speech. Furthermore, the rules moderators enforce vary widely both in their formulation and interpretation across communities, making a one-size-fits-all approach increasingly brittle. Since successful moderation relies on fine-grained understanding of a given community's norms, we present a new dataset and models for community-specific, contextualized norm violation detection for over twenty types of norms.

[1]Dataset, code, and models are publicly available at https://github.com/chan0park/NormVio.

[2]https://backlinko.com/reddit-users#reddit-statistics

[3]All example comments used in this paper are lightly paraphrased to preserve privacy.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3386–3397
November 7–11, 2021.

We introduce a new approach to context-sensitive automated content moderation that explicitly encodes community norms. Using a new dataset of 51K conversations across 3.2K communities, we show that the most commonly-studied norm violation behavior in NLP, hate speech, corresponds to a small minority of cases in which moderators intervene in practice. We then create multiple models to detect when moderators intervene and *why* they intervene, adapting to the norms and rules of a community.

Our paper offers the following four contributions towards advancing the future of NLP in community and context-specific moderation. First, in a large scale analysis of rule and moderation behavior, we show that subreddits vary considerably in their rules, with only some common themes. However, in practice, most rules are not enforced and, further, the enforcement of some types of rules, e.g., *incivility*, is highly varied across communities. Second, we introduce a new dataset, NORMVIO, of 51K conversations across 3.2K subreddits and map the 25K rules from these communities into nine categories of context-specific unacceptable behavior, including five types of incivility. Third, we introduce a new series of models aimed at detecting and explaining rule-violating behavior based on norms and rules of the community. Our approach enables not only identifying that conversation in a particular community (with particular rules) is likely to violate a rule, but also *which rule*. We demonstrate the effectiveness of these models, showing our best model attains an F1 of 78.64 across all rule types, a 50% improvement over context-insensitive baselines. Finally, we perform an in-depth analysis of how much conversation context and community-sensitivity affects predictability. Our work points towards key challenges in detecting particular rule violations, while providing high accuracy in others, which can allow moderators to quickly intervene. More generally, our work provides a clear next step for NLP to look beyond one-size-fits-all methods for detecting incivility to developing holistic, context-sensitive approaches that better suit the needs of moderators and their communities.

2 NORMVIO Dataset

Prior work has created datasets used to detect single types of norm violations in social media messages (e.g. incivility, hate speech or hostility) (Waseem and Hovy, 2016; Founta et al., 2018). However, these datasets typically focus on isolated texts and do not provide prior conversational context or community-specific details.

In order to detect representative types of norm violations and account for context, we construct a new dataset—NORMVIO—a collection of 52K English conversation threads on Reddit. NORMVIO includes comments removed for violating a variety of community norms beyond the traditional hate speech and incivility, such as spamming or violating community format/topics. Furthermore, NORMVIO provides additional context beyond the norm-violating comment itself with (a) the entire conversation thread (i.e., the original post and prior comments) and (b) the subreddit (i.e., community) in which the comment was posted.

Data Collection We collected our initial data via the Reddit API, which provides list of moderators and their comments for each subreddit. For each of the top 100K most popular subreddits,[4] we identified the most recent 500 comments from each moderator and retrieved comments that moderators posted in response to a removed comment (henceforth, *moderation comments*).

Moderation comments often provide useful signals for inferring which community norm was violated. From the full set of moderation comments, we selected those that contain a phrase explicitly stating the rule number (e.g. "this comment violates Rule 2") or the exact text of one of its subreddit's rules (e.g. "don't be rude").

We then fetch the entire conversation thread for this set of moderation comments: the original post and all parent comments prior to the moderator's comment. We also fetched the norm-violating comment that was removed by moderators, by searching archived comments via the Pushshift API (Baumgartner et al., 2020).[5]

The final dataset is comprised of 20K conversations that have the last comment removed by one of the moderators of the community. Following the approach in Chang and Danescu-Niculescu-Mizil (2019), we include 32K paired unmoderated conversations as a control set. Each moderated conversation is matched with up to two unmoderated conversations from the same post and with most

[4]Ranked by number of subscribers as of April 2021

[5]We were unable to retrieve an additional 21K removed norm-violating comments, which were unavailable in the PushShift archive. We still include these corresponding conversations in our data release as they can be useful in the task of forecasting future norm violations.

similar conversation lengths as the target moderated conversation.

Ethical Considerations for Protecting User Privacy Our dataset focuses, in part, on comments that moderators have viewed as objectionable and therefore removed. While these moderated comments are still publicly available, their use requires additional ethical reflection and precautions to preserve the dignity and privacy of users (Townsend and Wallace, 2016). Moderated comments offer significant benefit to the study of supporting moderators and authorities in their goals of having supportive technologies that match their community's norms. At the same time, users who made those comments may object to having them included in a dataset (Fiesler and Proferes, 2018). Therefore, we take additional measures to ensure that user privacy is protected, especially for the deleted comments.

We use Reddit data through Pushshift (Baumgartner et al., 2020), an archive that has been widely used in NLP and related fields since its first release in 2015 (Hessel and Lee, 2019; Kennedy et al., 2020; Sap et al., 2020; Dinan et al., 2020, *among many others*). Pushshift's collection policy explicitly states that it conforms to Reddit's rules and user agreement with regards to data collection. In releasing our dataset, we provide only the associated identifiers of comments but *not* their textual content. Practitioners will need to independently fetch the texts from Pushshift by using the provided comment IDs. Releasing only IDs ensures that any users who request their data to be removed in Pushshift will also have it removed in our dataset. Additionally, in our dataset we anonymize individual usernames and personal identifiers of posters and moderators. Finally, along with our data release, we provide guidelines to the users who wish to delete their comments from the Pushshift dump.

Classification of Community Norms Moderator comments as well as rules defined in each subreddit are free-form and diverse, and it is not trivial to map the rule/comment to a specific community norm it refers to. In order to study norm violations, we thus first train classifiers that given a rule description label it with a type of norm it violates.

We follow Fiesler et al. (2018)'s qualitative analysis of 1K subreddits, that identified main categories of rules through annotating 3,789 rules from the subreddits.[6] We then use the annotations from (Fiesler et al., 2018) to fine-tune a BERT-based binary classifier for each rule type.[7] Table 1 shows the list of the resulting 21 categories of community norms and the performance of our classifiers evaluated using macro F1 scores with stratified 10-fold cross validation.

Rule Types	F1	Rule Types	F1
Advertising	71.0	NSFW	88.2
Moderation Enforcement	87.0	Off-topic	63.5
Copyright/Piracy	70.6	Personal Army	43.2
Doxxing	75.4	Personality	81.9
Format	73.5	Politics	85.7
Harassment	67.9	Reddiquette	83.2
Hate Speech	84.2	Reposting	81.4
Images	65.1	Spam	86.9
Outside Content	68.0	Spoilers	76.7
Low-Quality Content	45.6	Trolling	96.0
		Voting	85.6
AVERAGE	75.3		

Table 1: Macro F1 of classifying the diverse sets of rules across subreddits to rule violation types.

We use the final models to map 183K rules from the top 100K subreddits to their corresponding rule types. Table 2 shows the examples of labeled community rules randomly sampled from our data. Finally, we classify moderators' explanations of the rule-violating comments in NORMVIO. Because we only kept moderators' comments that mention a rule number or a rule's exact text, we can determine which rule was violated by the conversation.[8] Using our binary classifiers on rule text, we can now infer the type of norm that was violated by the moderated (removed) comment.

Although the 21 types are well suited for fine-grained analysis of rules on Reddit, they might leave insufficient number of examples per type which can make it more challenging to compu-

[6] Out of 24 categories, we exclude the ones describing the tone of rules (whether a rule is "Prescriptive" or "Restrictive") and one (Behavior/Content) that is extremely broad, covering over 90% of coded rules.

[7] Binary classifiers were used since each community rule can be categorized with multiple types. We used the default hyperparameters suggested in the Transformers library and trained each model for 20 epochs.

[8] Any data collection procedure that relies on user-generated labels has the risk to absorb human biases. In our setting too, there is a risk of moderator biases to be incorporated when we match moderation comments to rules and violation types. However, in pilot work examining moderator comments with explicit rule violations and those where we had to infer the rule(s), we found a near-identical distribution of violation types.

Incivility: {Personality}	*"Be civil"*
Harassment: {Harassment, Doxxing}	*"Don't harass others"*
Spam:{Spam, Reposting, Copyright}	*"No excessive posting"*
Format: {Format, Images, Links}	*"Use the correct tags"*
Content:{Low-quality Content, NSFW, Spoilers}	*"No low-quality posts"*
Off-topic :{Off-topic, Politics}	*"Only relevant posts"*
Hate speech:{Hatespeech}	*"No racism, sexism"*
Trolling:{Trolling, Personal Army}	*"No trolls or bots"*
Meta-rules:{Voting, Moderation Enforcement, Reddiquette}	*"No Downvoting"*

Table 2: The mapping between coarse- and fine-grained rule types and examples.

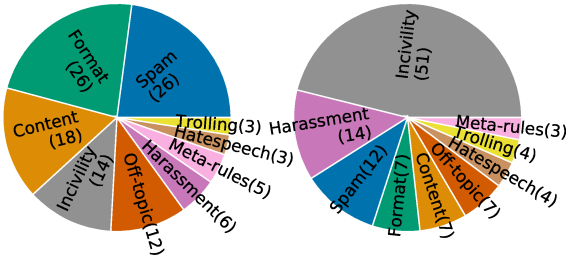

Figure 2: % of rule types of rules (left) and comments violating those rules (right) in NORMVIO.

tationally model them. We define relatively more coarse-grained nine types and map the 21 types with the nine types as shown in Table 2. We designed these types to reflect our interest in text-based analysis of abusive language. We kept five different subcategories of uncivil comments (general incivility, trolling, harassment, hate speech, spam) while aggregating Voting, Reddiquette, and Moderation Enforcement into a broad "Meta-rules" category. In the remainder of this paper, we only use the coarse-level norm violation types.

Ultimately, each moderated comment in NORMVIO has the following information: (1) its subreddit, (2) its conversation thread, (3) the community-specific rule violated, and (4) the coarse- and fine-grained rule types that were violated. To maximize user privacy, all comments are provided as IDs, the content for which can be retrieved through the Reddit and PushShift APIs.

Analysis of Community Norm Violations We analyze the types of rules and comments comprising NORMVIO with a focus on what kinds of rules are established by communities, what kinds of rules are violated in practice, and when in conversations these rules are violated.

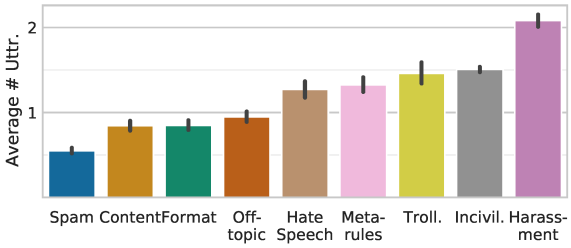

Figure 3: Average number of utterances between the original post and the moderated comments.

The results in Figure 2 show that the rule types are evenly distributed over rules (left) while the actual violations (right) are relatively more focused on abusive language rule types such as Incivility and Harassment. A large proportion of all rules in our dataset fall under the Format and Content categories, suggesting that there is a diverse set of community norms, beyond regulating incivility, needed to operate healthy online communities. Critically, while the majority of efforts on identifying abusive language in the NLP community have been focused on hate speech, more subtle types of incivility are significantly more prevalent in removed comments, which are also harder to detect (Jurgens et al., 2019; Breitfeller et al., 2019; Field and Tsvetkov, 2020). Moreover, only 55% of removed comments are violations of Incivility and Hate Speech rules, again highlighting the importance of understanding the spectrum of community norms in designing automated moderation assistance systems.

Figure 3 shows the average number of utterances from the original post to the norm-violating removed comment. Overall, violations related to abusive language such as Harassment, Incivility, and Trolling occur *later* in conversations than comments removed for other reasons (e.g. Spam and Format). This timing has implications for the "forecastability" of violation types. For example, the average conversation length within the Spam category is about 0.5 which indicates that half of the violations happen in the original post or a reply to it, making it impractical trying to forecast such violations.

Even though Hate Speech and Harassment are both related to abusive language, comments removed due to Harassment occur after more inter-

actions. We hypothesize this is because harassment and trolling are intentionally expressed in less overt forms to delay the moderators' intervention. These findings illustrate that with a more representative set of community rules and a larger-scale dataset, NORMVIO facilitates deeper understanding of community norm violation behaviors and provides guidance on more urgent tasks our field should be focusing on for a practical impact.

3 Detecting and Explaining Community Norm Violations

With NORMVIO, we can now train models for detecting contextualized, fine-grained community norm violations. We present two tasks: (1) Detecting community norm violations, and (2) Explaining community norm violations. The former identifies coarse categories of norm violations detailed in §2, and the latter is aimed at identifying specific local community rules being violated, to facilitate moderation transparency. For each task, we compare model variants without or with varying types of incorporated context, including conversation history and community information (e.g. subreddit name).

3.1 Detecting Community Norm Violations

In this task we assume a set of pre-defined categories of norm violations. For each category, we train a binary classifier to detect violations, since the categories are not mutually exclusive.

As shown in Figure 4, we encode a conversational context of arbitrary length along with community rules. Following Chang and Danescu-Niculescu-Mizil (2019), we use a uni-directional LSTM context encoder. The utterance encoder is initialized with a pretrained BERT model, with each classifier is then fine-tuned using training data specific to each rule type (data statistics are detailed in Appendix A). The last hidden state from the last comment is fed into the classifier. The flexibility of this design allows for both retroactive detection after violations occur (the focus of this work) as well as proactive prediction of future rule violations.

We experiment with four model variants with different input contexts:

- COMMENT: Only the final comment.
- +HISTORY: Past conversation history and the final comment.
- +COMMUNITY: Community information and the final comment. We concatenated the subreddit name in front of the comment (e.g. "r/AskReddit ask anything!").[9]
- +HISTORY+COMMUNITY: Conversation history and community information.

3.2 Explaining Community Rule Violations

In addition to categorizing rule violations by type (type-based), we develop a model that leverages the specific community rule text to identify violations in context. This text-based model facilitates explanations of rule violations, and improves transparency (Juneja et al., 2020). Such a system could lighten moderators' workload through highlighting why they might moderate a comment, enable more productive interventions, and improve the relationship between community members and moderators.

Similar to the violation category detection task, we construct binary classifiers that detect violations given conversational and community context. However, as shown in Figure 4, the full input and training procedure are different; we include the community's verbatim rule description as a model input. The rule text is appended to the input comment with a special token ([SEP]) added between the comment and the rule to leverage pretrained language models' ability to infer relationships between two sentences. Since the precise formulation of the target rule is given as an input, we no longer need to train one model per rule type; we train one universal model with all available training data.

NORMVIO contains information about which rules are violated in each removed comment, and we use these rule-comment pairs as positive examples. If a comment is tagged for violating more than one rule, we include all comment-rule pairs as positive examples. We construct negative training examples using matched unmoderated conversations from NORMVIO (described in §2) by adding the text of the violated rule to the corresponding unmoderated conversation.

To guide the model in better discriminating rules, we construct additional negative examples by mapping each removed comment with an randomly chosen incorrect rule from the same subreddit (e.g. "Here's my referral code! [SEP] No Politics").

Similarly, we experiment with three model variants with different input contexts:

- +RULE: Only the final comment and a rule text.
- +RULE+HISTORY: Past conversation history,

[9] Note that the model variants without conversation history do not use a context encoder at all and thus have a smaller number of trainable parameters.

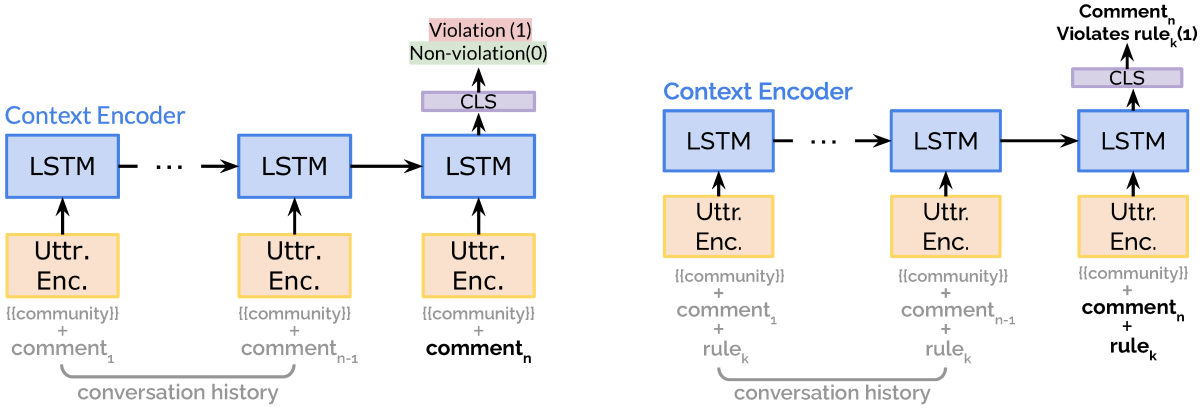

Figure 4: Structure of the baselines of the two proposed tasks: detecting norm violation (left) and explaining norm violation (right). Inputs in gray (conversation history and community information) are optional context.

the final comment, and a rule text.

- +RULE+HISTORY+COMMUNITY: Both conversation and community history, the final comment, and a rule text.

The main advantage of the text-based model is in its interpretability and generalizability. Since the model now looks at the community-specific rule texts, the system can provide more meaningful feedback to moderators and users. For example, instead of saying "potential hate speech detected", now the model can be more informative in notifying users that "the comment has breached our community's Rule 2: No Racial Slurs". Moreover, since the model takes free-form rules as input, it can generalize to unseen rules and novel rule types,

4 Experiments

Baselines In addition to the seven model variants in §3, we consider three baselines that represent current common approaches:

- MAJORITY: Majority class baseline.
- PERSPECTIVE: Perspective API's toxicity score of the final comment to make a binary decision. For each rule type, a threshold value was tuned to maximize development set F1 score.
- INCIVILHATE: We train a model using just the incivility and hate speech violations from NORMVIO. The test set predictions from the trained model was evaluated over different rule types.

Training Details We perform an 80-10-10 train/dev/test random split of moderated comments in NORMVIO and then appended paired unmoderated comments into the same split. The resulting number of examples of train/dev/test split was 41667, 5214, and 5131, respectively. We ran training for five different random seeds and report the average scores of multiple runs except for MAJORITY and PERSPECTIVE baselines.

The base utterance encoder is a pretrained Conversational BERT model. Each model was trained for 10 epochs with an early stopping patience of 5, and with Adam optimizer with a learning rate of 1e-5. We used a batch size of 32 for models that do not leverage past conversations and 8 for the ones that use comment history. We used 2 layers of GRUs with a hidden size of 768 for the context encoder and 2 linear layers for the final classifier.

Evaluation We used macro F1 to evaluate all models. For models in §3.2, at test time we cannot assume that we know which rule will be violated in a given conversation. We thus create multiple comment-rule pairs for each comment in the test set by matching it with each community rule. Out of the resulting pairs, we mark the pairs that were observed in the original test set as positive, and the remaining pairs are marked as negative. We refer to these negative pairs added to the test set of models explaining rule violations as *augmented pairs*. Note that the test sets of models in §3.2 are now different from the text sets in §3.1 and the F1 scores of two tasks are not directly comparable.

Experiment Results Information from the social context of a comment substantially improves performance (Figure 5). Compared to current approaches for inferring toxicity, all type-based violation detection model performed significantly better—even for rule violation categories those approaches are tailored for. While PERSPECTIVE

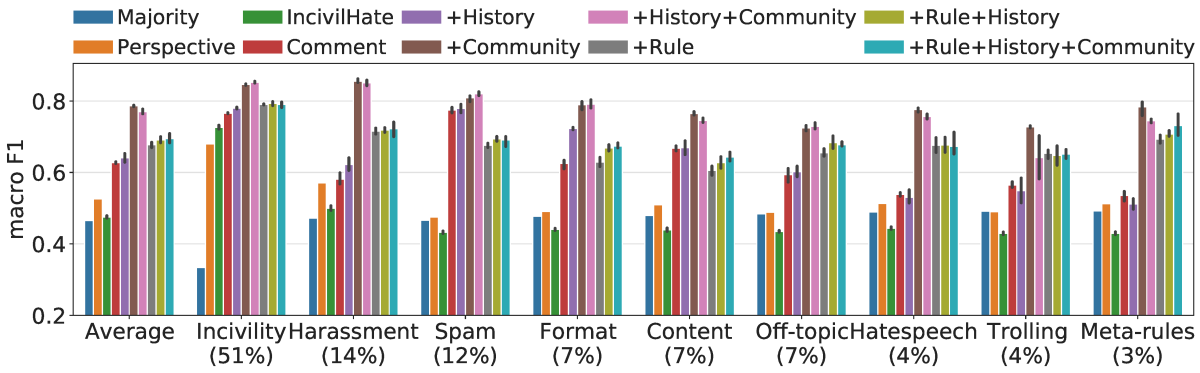

Figure 5: Average and breakdown of Macro F1 scores of the baselines and the model variants. Error bars indicate 95% confidence interval and the types are sorted by their violation frequency (percentage below x-axis labels).

and INCIVILHATE performed better in Incivility and almost comparable with COMMENT and +HISTORY, adding community information still resulted in a significant improvement of +8.0 absolute increase in F1.

Across all rule violation types, adding the context about community significantly improved the performance, often resulting in the highest performing models when added. Adding conversation history showed mixed results. +HISTORY showed improvements over COMMENT whereas +HISTORY+COMMUNITY was not necessarily better than +COMMUNITY. Models with conversation history tend to perform worse on scarce violation types such as Meta-rules and Trolling; we speculate that this decreased performance is due to the increased number of parameters from adding context encoder layer to process conversation history and future work with more examples of these violations may substantially improve performance. This result for history greatly expands an analysis by Pavlopoulos et al. (2020) that found minimal performance gain when adding a single prior comment to identify toxicity; while we too find minimal improvement for Incivility and Harassment norms, adding history *does* improve the recognition for other norm violations (e.g., Format and Content) indicating that prior context can be useful.

While the results of text-based violation detection models (+RULE, +RULE+HISTORY, +RULE+HISTORY+COMMUNITY) and type-based models are not directly comparable due to the augmented pairs, they were evaluated over the same set of comments so the numbers can provide a general sense of text-based model performance. An interesting distinction between the two detection tasks is in how much additional context helps. In type-based models, adding context made significant improvements in all or in some cases. However, with text-based models, the performance was relatively more uniform and additional context did not contribute as much. This result suggests that providing full text of rules may help resolve certain ambiguous comments and thus the model rely less on the additional context.

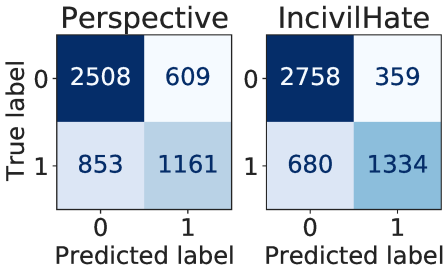

Figure 6: Confusion matrices of two baselines over the norm violation detection task.

5 Analysis

How many violations do current systems miss? In part due to their targeted focus, the PERSPECTIVE and INCIVILHATE baseline models miss a substantial proportion of the total norm violations. Figure 6 shows the confusion matrices of the violation detection task, where labels are aggregated over all violation types to test how many violations overall are not captured by these systems. The results show that PERSPECTIVE and INCIVILHATE fail to recognize 42% and 34% of all violations, respectively. Moderators on platforms like Reddit must triage huge numbers of comments daily and this points to a clear gap between current practice (represented by the baselines) and indicates what moderators act on in practice.

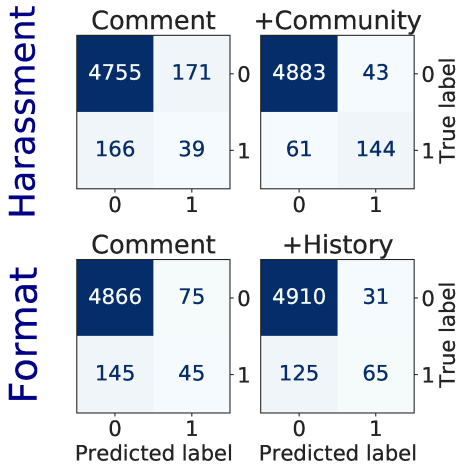

Figure 7: Confusion matrices of COMMENT and +COMMUNITY for the Harassment detection task (top), and of COMMENT and +HISTORY for the Format violation detection task (bottom).

How does community information help? We observed that adding community information provides the most significant improvements in Harassment in Figure 5. We now look into the Harassment type to understand more about how did the additional community information actually help to improve the performance.

What kinds of errors are corrected by adding community context? By comparing confusion matrices of COMMENT and +COMMUNITY (Figure 7), we find that +COMMUNITY has fewer false positives. Out of 154 false positives from the COMMENT model that were corrected in the +COMMUNITY model, 106 (69%) were Incivility violations. Consider the following example:

Comment:
That game's already dead to 99% of the world a few weeks later, get over it you stupid idiot.
Moderator Comment:
Your comment has been removed for Rule 2. Be civil and respectful. Do not attack or harass other users or engage in hate-speech.
Paired Rule: Rule 2: Be civil and respectful.
Violation: Incivility
Community: r/classicwow

The final comment in this example could be considered as both a Incivility and Harassment violation and COMMENT model labels it as Harassment. Although the moderator refers to the community's Incivility rule, the rule mentions "do not attack or *harass* other users", which makes it clear that this example falls into both categories. However, the +COMMUNITY model labels this comment as Incivility and not Harassment. We speculate that the +COMMUNITY model learns about what rules exist in each community; *r/classicwow* has 8 rules and none of them are about Harassment, so moderators refer to the Incivility rule when moderating Harassment violations. In other words, depending on the community and their available community rules, the same comments can be moderated as either incivility or harassment violation. Therefore, providing the community information can help the model disambiguate this decision and ground its moderator support in the norms of the community.

How does conversation history help? Likewise, for the conversation history context, the largest gain was achieved in the Format type. In Figure 7, we compare confusion matrices of COMMENT and +HISTORY. The result again shows that additional context can help the model in reducing the false positive rate.

Among the corrected false positives, the most prevalent type mistaken for Format was Spam. One example of such case is given below:

Comment: *UPDATE: I found it! here you go if you need it_LINK_*
Violation: Spam (Piracy)
Moderator Comment:
See Rule 1: No Merchandise / Spam
Previous Message:
Does anyone know where to buy this?

If we only consider the final comment, there are two possible explanations for which rule was violated: 1) Format: the outside link does not follow the community guideline 2) Spam: self-promotion / promoting specific merchandise is banned. However, the previous message makes it clear that the author had just posted about a product and then made a self-reply with a link to buy the product. With this information, model can disambiguate this situation and choose the right violation type.

6 Related Work

Community Norms and Rules Many studies have investigated how online conversations are moderated and how each community has different norms to ensure a safe environment for discussions (Chandrasekharan et al., 2018; Jhaver et al., 2018, 2019; Juneja et al., 2020; Almerekhi et al., 2020;

Rajadesingan et al., 2020). Fiesler et al. (2018) conduct an analysis over the rules of Reddit communities and define 24 types of the rules. They provide a thorough and large-scale analysis over how the rules are phrased and how rules are different across subreddits. We adopt their rule categorization and extend it to code actual rule violations.

Chandrasekharan et al. (2018) also studied removed comments on Reddit to understand what types of rules exist on Reddit by clustering the moderator comments and investigated how they are governed. However, their dataset provides limited context of moderated comments, whereas we focus on providing a dataset that has enough context and also explicit violation type that can be leveraged in modeling rule violation.

Context in Detecting Online Abuse Most of the existing datasets for abusive language detection implicitly assumes that comments may be judged independently taken out of context. Pavlopoulos et al. (2020) challenged this assumption and examined if context matters in toxic language detection. While they found a significant number of human annotation labels were changed when context is additionally given, they could not find evidence that context actually improves the performance of classifiers. Our work also examines the importance of context, but we do not limit our scope to toxic language detection and investigate a broader set of community norm violation ranging from formatting issues to trolling.

Beyond Incivility and Hate Speech Jurgens et al. (2019) claims "abusive behavior online falls along a spectrum, and current approaches focus only on a narrow range" and urges to expand the scope of problems in online abuse. Most work on online conversation has been focused on certain types of rule violation such as incivility and toxic language (e.g., Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019; Almerekhi et al., 2020). In this work, we focus on a broader concept of *community norm violation* and provide a new dataset and tasks to facilitate future research in this direction.

7 Conclusion

Online communities establish their own norms for what is acceptable behavior. However, current NLP methods for identifying unacceptable behavior have largely overlooked the context in which comments are made, and, moreover, have focused on a relatively small set of unacceptable behaviors such as incivility. In this work, we introduce a new dataset, NORMVIO, of 51K conversations grounded with community-specific judgements of which rule is violated. Using this data, we develop new models for detecting context-sensitive rule violations, demonstrating that across nine categories of rules, by incorporating community and conversation history as context, our best model provides a nearly 50% improvement over context-insensitive baselines; further, we show that using our models, we can *explain* which rule is violated, providing a key assistive technology for helping moderators identify content not appropriate to their specific community and better communicate to users why. Our work enables a critical new direction for NLP to develop holistic, context-sensitive approaches that support the needs of moderators and communities.

8 Ethical Considerations

We hope to draw attention to the mismatch between the standard tasks of harmful content detection that NLP researchers are typically focusing on (e.g. sentence-level toxicity detection) and the broad spectrum of context-sensitive content violation types that actually occur in the wild. To enable future research on detecting community-specific norm violations, we constructed a dataset that retrieves online conversation threads and comments deleted by moderators, categorized by community norm violations. We discuss ethical considerations related to protecting user privacy in §2.

Additionally, we acknowledge that the dataset itself can incorporate unintentional biases. For example, it can incorporate moderators' biases in deciding which comments are selected to be removed (Binns et al., 2017; Myers West, 2018; Shen and Rose, 2019). The unmoderated comments can include norm-violating comments that were missed by the moderators (Chandrasekharan et al., 2018). By constructing a large scale dataset that spans multiple subreddits and moderators' teams we partially mitigate these concerns. To investigate this further, future work could incorporate an additional evaluation procedure with test sets containing held-out moderators (cf. Geva et al., 2019).

Acknowledgments

This material is based upon work funded by the DARPA CMO under Contract No. HR001120C0124, and by the National Science Foundation under Grants No. IIS2040926 and 1850221. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, pages 294–298.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. 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 4743–4754, Hong Kong, China. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 307–318. International World Wide Web Conferences Steering Committee / ACM.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Bryan Dosono and Bryan C. Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 142. ACM.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of Twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.

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.

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

Jack Hessel and Lillian Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. 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 1648–1659, Minneapolis, Minnesota. Association for Computational Linguistics.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33.

Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

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

Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.

Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "eternal september": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 1152–1156. ACM.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *University of Aberdeen*, 1:16.

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, San Diego, California. Association for Computational Linguistics.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Dataset Description

NORMVIO	
# of total comments	52012
# of moderated comments with original final comment restored	20137
# of unmoderated comments	31875
Additional dataset for forecasting (without original violation comments)	
# of total comments	53829
# of moderated comments	20727
# of unmoderated comments	33102
# of subreddits	3234
# of rules	24916
# of moderators	29841
# of moderators per subreddit	9.2
Avg. comment length (# of words)	34.4
Avg. number of context per comment (including the original post)	2.8
Avg. # of rules per community	7.7

Table 3: Summary statistics of NORMVIO

Table 3 presents the basic summary statistics of NORMVIO. Our main dataset used in the analysis consist of 52K comments in total, and each comment is accompanied with its conversation history, subreddit information, tagged rule, and its violation type.

The dataset also provides additional 54K comments that contains 21K violation comments and its paired 33K unmoderated comments. For these moderated comments, we could not fetch its original comment before getting moderated, so these could not be used for detection task. However, these comments could still be used in training norm violation *forecasting* models.

B Additional Details for Reproducibility

Our work includes two series of model training: rule classifier training and violation detection model training. For all training runs we trained with one GPU with 11GB of memory.

For rule classifiers, we had to train one binary model for each violation type, so we had to run 21 final training using 3.7K annotated rules. Each run took about 5-6 minutes which results in about 2 hours of training.

Violation detection models are trained with 52K examples thus took significantly longer than training rule classifiers. Again, for type-based detection models, we needed to train one model per coarse-grained violation types. Each run took about 40 minutes for models without conversation history and took about 2 hours for models with history. In summary, to run one set of training for one model, we needed to train for 6 hours for models without history and 18 hours for models with history.

For text-based detection models, we did not need to train a model per type which significantly reduces the total training amount. Models without conversation history took about an hour to train and models with history took about 7 hours to train one model.

The number of trainable parameters was 109 million for models without conversation history (i.e., those without a context encoder) and 116 million for models with a context encoder.