The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology, USA
MATTIA SAMORY, Virginia Tech, USA
SHAGUN JHAVER, Georgia Institute of Technology, USA
HUNTER CHARVAT, University of Michigan, USA
AMY BRUCKMAN, Georgia Institute of Technology, USA
CLIFF LAMPE, University of Michigan, USA
JACOB EISENSTEIN, Georgia Institute of Technology, USA
ERIC GILBERT, University of Michigan, USA

Norms are central to how online communities are governed. Yet, norms are also emergent, arise from interaction, and can vary significantly between communities—making them challenging to study at scale. In this paper, we study community norms on Reddit in a large-scale, empirical manner. Via 2.8M comments removed by moderators of 100 top subreddits over 10 months, we use both computational and qualitative methods to identify three types of norms: *macro* norms that are universal to most parts of Reddit; *meso* norms that are shared across certain groups of subreddits; and *micro* norms that are specific to individual, relatively unique subreddits. Given the size of Reddit's user base—and the wide range of topics covered by different subreddits—we argue this represents the first large-scale study of norms across disparate online communities. In other words, these findings shed light on what Reddit values, and how widely-held those values are. We conclude by discussing implications for the design of new and existing online communities.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: online communities; community norms; moderation; mixed methods.

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (November 2018), 25 pages. https://doi.org/10.1145/3274301

1 INTRODUCTION

An online community's norms play an important role in guiding acceptable behaviors, and therefore in its governance [29]. Online community moderators have to sanction pedestrian normative

Authors' addresses: Eshwar Chandrasekharan, Georgia Institute of Technology, School of Interactive Computing, USA, eshwar3@gatech.edu; Mattia Samory, Virginia Tech, Department of Computer Science, USA, samory@vt.edu; Shagun Jhaver, Georgia Institute of Technology, School of Interactive Computing, USA, jhaver.shagun@gatech.edu; Hunter Charvat, University of Michigan, School of Information, USA, charvat@umich.edu; Amy Bruckman, Georgia Institute of Technology, School of Interactive Computing, USA, asb@cc.gatech.edu; Cliff Lampe, University of Michigan, School of Information, USA, cacl@umich.edu; Jacob Eisenstein, Georgia Institute of Technology, School of Interactive Computing, USA, jacobe@gatech.edu; Eric Gilbert, University of Michigan, School of Information, USA, eegg@umich.edu.

2573-0142/2018/11-ART32 $15.00
https://doi.org/10.1145/3274301

violations like posting spoilers about a TV show, as well as more serious infractions like online abuse [39], harassment [8, 17, 18], and fake news and misinformation [41]. Yet, norms for what is appropriate can vary widely from one community to another. Even behavior considered harmful in one community might be celebrated in another (e.g., 4chan's /b/ [1], Something Awful Forums [34]).

1.1 Regulating behavior on Reddit

In this paper, we study norms across a wide variety of communities on Reddit. Reddit is an assemblage of over one million online communities[1] known as *subreddits*. Subreddits can be created by anyone, and they are moderated by members of the community. They exist for almost any topic, including specific sports (e.g., r/nba), science (e.g., r/science), TV fan theories (e.g., r/gameofthrones), and standing cats (e.g., r/standingcats).

Reddit has a multi-layered architecture for regulating behavior on the platform. It has site-wide content[2] and anti-harassment[3] policies that all subreddits are expected to follow. In cases where there are violations of some of these policies, Reddit is known to ban subreddits and user accounts [8]. In addition to Reddit's content policies, each subreddit has its own set of subreddit-specific rules and guidelines regarding submissions, comments, and user behaviors [19]. Moderators (or "mods") enforce the rules and guidelines.

1.2 Community norms on Reddit

Rules and norms are loosely coupled on Reddit, with subreddit moderators sometimes turning (often implicit) norms that are enforced behind-the-scenes into rules that face the community. While rules tend to be explicit, norms are emergent, arise from interaction over time, and respond to current demands on a community [33]. Community norms play an important role in online moderation, and moderating online communities is strongly contextual because norms can vary widely between communities. An understanding of community norms is generally gained through experience [4]: observing posts and comments posted on the subreddit, peer feedback in the form of votes or replies to comments, and interactions with mods. This work of enforcing norms is important to both communities and platforms: people may leave sites and communities after being the victims of norm violations [26]. Importantly for the present work, norm enforcement by mods also creates a record of norm violations across disparate communities.

1.3 Summary of methods, findings and contributions

In this paper, we study community norms on Reddit with a large-scale, empirical approach. By working from over 2.8M comments removed by moderators of 100 subreddits over 10 months, we use both computational and qualitative methods to identify three types of norms within Reddit: *macro* norms that are universal to most parts of Reddit; *meso* norms that are shared across certain large groups of subreddits; and *micro* norms that are specific to individual, relatively unique subreddits.

1.3.1 Summary of methods. A flowchart describing all the components of our research pipeline is shown in Figure 1. We first train linguistic classifiers for 100 top subreddits, using moderator-removed comments from each subreddit; those classifiers only "see" their own subreddit's data and predict moderator removals in that subreddit. Next, we ask the classifiers to estimate a counterfactual: For every comment in our dataset, *what would this subreddit have done if this comment had been posted there?* Using this, we cluster subreddits that often agree to remove the same comments (based on their classifiers' predictions). Finally, we compile all comments that subreddits within

[1] http://redditmetrics.com/history
[2] https://www.redditinc.com/policies/content-policy
[3] https://redditblog.com/2015/05/14/promote-ideas-protect-people/

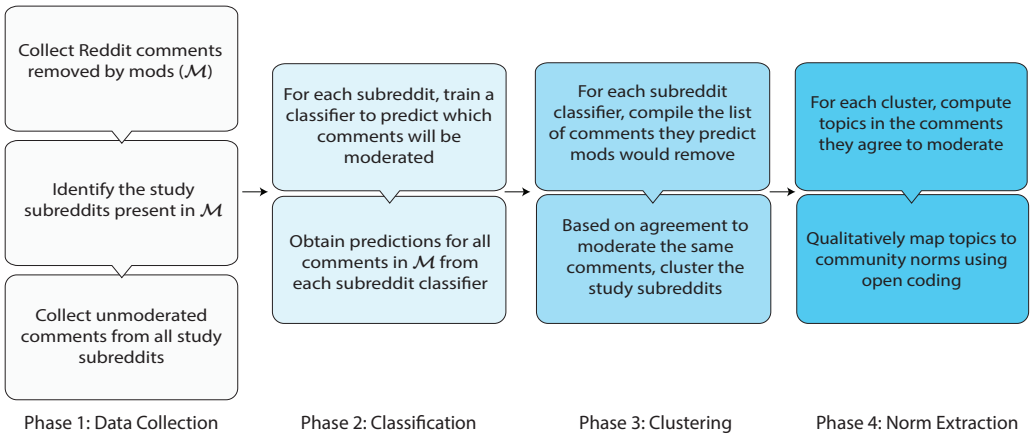

Fig. 1. Flowchart depicting the different phases of our research pipeline. $\mathcal{M}$ denotes all the moderated Reddit comments we collect in *Phase 1*, and *mods* denote the subreddit moderators on Reddit. The final output derived from *Phase 4* gives us the different community norms on Reddit.

each cluster agree to remove, and employ open coding to identify three different types of norms on Reddit: *macro*, *meso*, and *micro* norms.

1.3.2 Findings. Macro norm violations include employing personal attacks, misogyny, and hate speech in the form of racism and homophobia. In addition, *controversial views around Donald Trump*[4] , and *criticizing moderators* are norm violations on most parts of Reddit. Meso norms, by contrast, are not universal, and are only enforced by subgroups of subreddits. As expected, not sharing personal anecdotes, and not posting links to promotional spam are meso norms. Perhaps surprisingly, comments only *expressing thanks*, or acknowledging a good point, are meso norm violations. Furthermore, we observe that "mansplaining," mocking religion and nationality, and hostility toward immigrants exist only at meso scales—they are not considered norm violations on all of Reddit. Finally, we find highly specific micro norms that apply to individual, relatively unique subreddits. These are not widely enforced on most other parts of Reddit; one example is using high school-level science to explain new scientific discoveries (e.g., on r/AskScience).

1.3.3 Contributions and implications. Given the size of Reddit's user base—and the wide range of topics covered by different subreddits—we believe this work is the first large-scale study of norms across disparate online communities. In other words, these findings shed light on *what Reddit values*, and how widely-held those values are. For the design of online communities, it may be possible to use the frame of macro, meso, and micro norms to derive normative guidelines for new online communities. That is, the norms identified in this work may serve as sensible defaults for a new online community. Some norms, however, are problematic (e.g., *do not criticize mods*, *do not express thanks*) and suggest challenges for designing large-scale discussion systems. Finally, the discovery of widely overlapping norms suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values—a direction we discuss more fully at the end of the paper.

[4]We have seen instances where some subreddits disallow posting about Donald Trump so as not to attract the attention of Trump-supporters elsewhere on Reddit.

2 BACKGROUND

Lawrence Lessig argued that in social interactions mediated by computers, there are four factors that can be used to shape behavior: markets, architecture, policy and norms [29]. Next, we survey related research in two of these areas: *online moderation*, and *social norms*, both offline and online.

2.1 Online moderation

There are a variety of different approaches for regulating behavior in online communities. In a comprehensive meta-analysis, Keisler et al. present ways to limit the damage that bad behavior causes when it occurs, and to limit the amount of bad behavior that a bad actor can perform [27]. Current online platforms tend to rely on a combination of policy and design for regulating behavior. Policies are posted to make clear what is allowed and what is not [19] and then technical tools are used and human workers employed to enforce those rules. Technical tools depend on the ability to either edit or delete content (including users) or to append new information to content to inform future users. Sites like Reddit, Stack Overflow, and Yik Yak use *distributed social moderation* [32, 35]. On these sites, the content is moderated through a voting mechanism where registered users up-vote or down-vote each submission or comment. Such voting determines how prominently any content is displayed on the site. This model allows the community to collectively decide its threshold for what content is acceptable and which issues need to be articulated and discussed.

Online communities like Facebook groups and subreddits also use *centralized moderation* [27]. In this model, a small number of users called *moderators*, who are usually regulars from within the community, manually remove posts and comments that violate community norms. Such communities usually specify the rules for posting content on their forums, and these rules guide the moderation. This model often employs automated tools to flag posts (for example, posts containing any of a list of pre-specified offensive words or violating formatting requirements) for review by the moderators. In some communities, the moderators also review posts that are flagged by regular users on the community. After review, the moderators either remove the content from the site if they find it inappropriate for their subreddit, or allow it appear on the subreddit otherwise. Research on automatic approaches to moderating online antisocial behavior has shown that textual cyberbullying [16, 42] and undesirable posting [6, 9, 11, 38] can be identified based on topic models, presence of insults and user behavior.

Although the approaches mentioned above are widely used, they suffer from shortcomings. The first two approaches require a great deal of human labor. Particularly, in the centralized moderation approach, a few moderators have to spend countless hours in order to maintain the community [5, 28]. While some kinds of distributed moderation can be effective [8, 36], it can also make things worse and serve as a potential trigger for deviant behaviors [7, 10]. The literature around automated moderation approaches lack empirical studies about the effectiveness of various abusive content moderation strategies. This is largely due to the fact that when a site employs a moderation approach that removes content from the internet, it is therefore no longer visible.

Despite there being several studies about online moderation and building computational tools to assist moderators, regulating bad behavior still remains a pressing challenge for online communities [22, 23]. Therefore, an understanding of what norms are actually being enforced by moderators is important. We build on this line of research by deriving norms from removals by human moderators.

2.2 Social norms online and offline

Social norms are rules and standards that are understood by members of a group, and that guide and/or constrain social behavior without the force of laws [14]. These norms emerge out of interaction with others; they may or may not be stated explicitly, and any sanctions for deviating

from them come from fellow members of the social group, not the legal system. Norms vary to the extent to which they are *injunctive*, prescribing the valued social behavior, versus *descriptive*, informing us about how others act in similar situations [12, 13]. In addition to commonly accepted rules of desirable behavior, norms include rules forbidding unacceptable social behaviors, such as taboos against incest or infanticide, and laws or standards for conduct established by a government or elected body [40]. Norms shape our behavior related to more quotidian activities as well, from how loudly one should speak on a cell phone in a public space, to what the appropriate dress is in different social situations.

Regulation through policies, rules, and guidelines is not always visible, with governance occurring at the level of informal norms instead. Prior work on governance in online communities suggests the importance of social norms in regulating behavior, yet we also know that the difficulty for newcomers learning norms can lead to high drop-out rates [15]. Norms on Reddit are nested. Some norms are adopted from the general social context, for example that pejorative adjectives indicate rudeness. Some norms are shared across the internet, like all caps being the equivalent to shouting. Some norms are Reddit-wide, while others exist in some subreddits, but not others.

2.2.1 Rules vs norms. Rules and norms are interrelated, differing in their degree of explicitness [14]. Certainly in the context of Reddit, rules and norms are loosely coupled, with some mods in some subreddits turning norms that are enforced behind-the-scenes into explicit rules that face the community. Recent work has surveyed outward-facing subreddit rules [19], finding the frequencies of different rule types across Reddit (though approximately half of all subreddits have no explicit rules at all). It may be fair to think of Reddit rules as the front-stage to the norm's back-stage; that is, a rule is a formalized norm, and a norm is an informal rule, with a fluid boundary between the two. For the purposes of the present work, we treat norms as the emergent themes in the record of mod removals, some of which may overlap with explicitly formalized subreddit rules (however, a far greater proportion do not; see Tables 3–5.)

In this work, we leverage the language used in comments removed by moderators to identify and understand community norms. By exploring where these norms overlap across communities, the present work is the first we are aware of to compile a large-scale study of norms across disparate online communities.

3 DATA

Next, we transition to our dataset construction, and describe the procedure we use to collect Reddit comments removed by moderators. An illustration of this approach is shown in Figure 2.

3.1 Moderated comments from Reddit ($\mathcal{M}$)

We construct a dataset that includes all Reddit comments that were moderated off-site[5] during a 10-month period, from May 2016 to March 2017, in a three-stage process.

3.1.1 Stage 1: Stream Reddit comments into master log file continuously. We use the Reddit streaming API [6] to crawl all comments as they are posted on Reddit on a continuous basis. These are all comments posted to r/all, which can be from any subreddit that is not "private," and chooses to post its content to r/all. As we keep streaming comments continuously, we store all of the data in a master log file.

[5] Therefore, these comments are no longer publicly visible on the internet.
[6] https://praw.readthedocs.io/en/latest

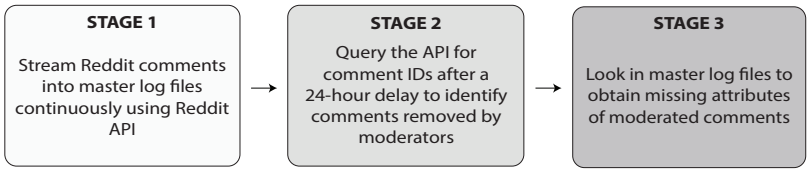

Fig. 2. Flowchart depicting the different stages involved in our collection of moderated (and unmoderated) comments from Reddit.

3.1.2 Stage 2: Query the API for comments after a 24-hour delay. After a 24-hour delay, we query the Reddit API for each comment in our master log file that was collected in the past day, using a comment's unique *comment_ID*. If a comment is removed by a moderator, then the text that was previously present in the comment (represented by the "body" field) is replaced with **["removed"]**.

Via conversations with Reddit moderators and an inspection of Reddit's source code[7], we know that only when moderators or admins remove a comment, its text is replaced by ["removed"], and most comments violating norms are moderated within the first 24 hours of posting on the subreddit. Note that a comment removed by a moderator (either auto- or human moderators) is different from an author removal, and we can distinguish between the two by looking at the text of the comment. The text in comments deleted by the authors is replaced by ["deleted"], while only moderator removals are replaced by ["removed"]. Using this method, we compile the "*comment_IDs*" of all comments that were removed by the moderators in the previous day.

3.1.3 Stage 3: Look in master log file to obtain missing attributes. For each moderated comment we identify in the previous stage, we perform a look-up in the master log file (compiled in Stage 1), using the *comment_IDs* of the removed comment. Through this look-up, we obtain all the fields that were previously contained in the removed comments (like 'body', 'subreddit', 'author', and so on) before it was removed by moderators.

Using this 3-stage process, we collect 4,605,947 moderated comments from Reddit during a 10-month period, from May 2016 to March 2017. All the moderated comments we identify constitute our Reddit moderated comments corpus (denoted by $\mathcal{M}$ in the remainder of the paper).

3.2 Preprocessing moderated comments in $\mathcal{M}$

3.2.1 AutoModerator replies. We observe the presence of comments authored by *AutoModerator* in our moderated comments corpus ($\mathcal{M}$). These are comments posted as replies to comments removed by the *AutoModerator*s of subreddits. *AutoModerator* is a customizable moderation bot used by many subreddits to automatically moderate posts from specific users or websites, and flag content that is inappropriate based on a predefined word list [8]. Upon removing a comment or link, *AutoModerator* posts a reply to the moderated comment indicating why it was removed. An example *Automoderator* comment is shown below:

> This submission has been removed. Submissions must be direct links to images in the imgur, minus, or gfycat domains. When using Imgur, simply right-click the image, select "Open in a new tab", and submit that URL. * I am a bot, and this action was performed

[7] Lines 1755-1766: https://github.com/reddit/reddit/blob/7471b22d90b39ef461769f082a17d6cbef1c9dff/r2/r2/models/link.py
Lines 738-746: https://github.com/reddit/reddit/blob/dbcf37afe2c5f5dd19f99b8a3484fc69eb27fcd5/r2/r2/lib/jsontemplates.py
[8] https://www.reddit.com/r/AutoModerator/

> automatically. Please [contact the moderators of this subreddit] if you have any questions or concerns.*

Given that different subreddits may use *AutoModerator* differently, we take precaution and remove all comments authored by *AutoModerator*, even if they do not appear in the form of replies to moderated comments in our dataset. Since these comments authored by *AutoModerator* are just warnings issued to users following actual removals, we do not consider them in our analysis. As a result, we discard all 101,502 comments which were authored by *AutoModerator* from $\mathcal{M}$.

3.2.2 Discarding replies to moderated comments. Next, we strip replies to moderated comments in $\mathcal{M}$. Through interactions with various subreddit moderators on a separate project, we learned that comments posted as replies to moderated comments are often also removed by moderators. These are replies that get removed due to their parent comment's removal, and they are sometimes referred to as the "children of the poisoned tree." Since these replies are not always removed intentionally, we decided to err on the side of caution, and discard such replies. We do this by identifying comments whose parents are themselves contained in $\mathcal{M}$. Through this procedure, we discard 1,051,623 moderated comments which we identify to be replies to comments that were removed, giving us the final dataset which we use for further analysis.

3.2.3 Study subreddits. After preprocessing the data, there are over 3 million moderated comments contained in $\mathcal{M}$, and they are collected from 41,097 unique subreddits. But we were only able to collect very few moderated comments (i.e., less than 10) for most of these subreddits. Our current goal is to build machine learning (ML) models that can predict moderator removals for the subreddits they are trained on. In order to build robust ML classifiers, we restrict our analysis only to the subreddits for which we were able to collect a reasonable amount of moderated comments. Therefore, we discard all subreddits that generate fewer than *5,000* moderated comments in $\mathcal{M}$.

Next, we discard all comments from any non-English subreddits present in our corpus. We use *langdetect* [9] and examine all comments from subreddits to decide whether the subreddit interactions are predominantly in English or not. For each subreddit, we predict only the top language using *langdetect*, and count the fraction of comments from that subreddit with English as first language. Via this step, we identify and discard 18 non-English subreddits, each of which contains more than 50% of their overall comments in languages other than English (e.g., r/podemos, r/svenskpolitik, r/Suomi, r/argentina, r/brasil, r/italy, r/france, and so on).

Finally, 2,831,664 moderated comments remain in $\mathcal{M}$, all originating in the 100 subreddits generating the most removed comments in our corpus. We call these 100 subreddits our *study subreddits* for the rest of the paper. At the time of our analysis, the study subreddits had an average 5.76 million subscribers, with r/funny having the highest subscriber count (19 million), and r/PurplePillDebate having the lowest subscriber count (16,000). On average, each subreddit contributes 20,070 moderated comments, with r/The_Donald contributing the most (184,168) and r/jailbreak contributing the least (5,616) number of moderated comments in $\mathcal{M}$.

3.3 Unmoderated comments from Reddit

In addition to collecting comments that were moderated from different subreddits, we also collect all comments that were not removed by moderators (i.e., unmoderated comments). As shown in Stage 1 of Figure 2, we store all comments obtained from r/all through the PRAW API in daily master log files. These master logs include comments that are both moderated subsequently after posting, and comments that still remained online at the time of data collection. In order to build our corpus of unmoderated comments, we use all comments present in the daily logs, which are

[9] https://pypi.python.org/pypi/langdetect

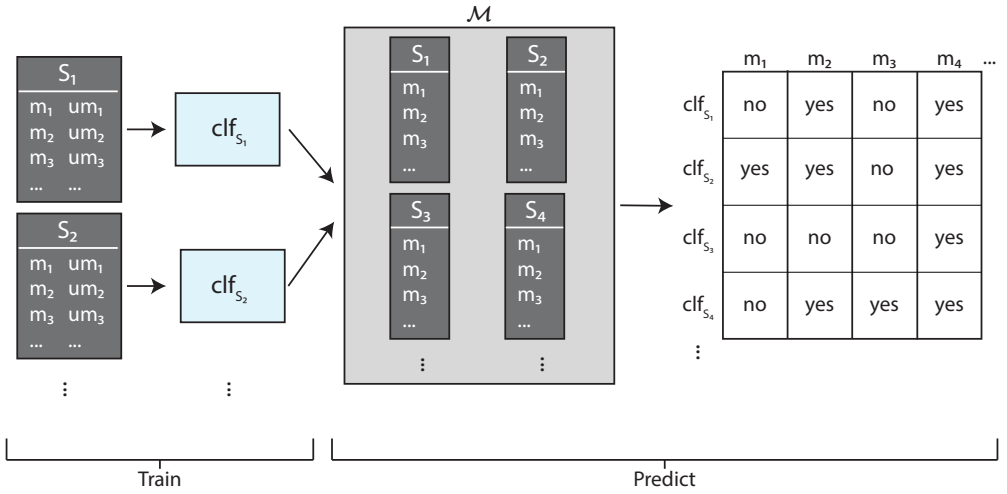

Fig. 3. In the first step (*Train*), we train classifiers to predict whether a comment posted on a subreddit will get moderated or not. For each study subreddit S_k, we build a classifier clf_{S_k} using moderated (e.g., m_i) and unmoderated (e.g., um_i) comments obtained entirely from S_k. In the next step (*Predict*), we obtain predictions from each subreddit classifier (e.g., clf_{S_k}) for each comment present in $\mathcal{M}$, and generate a *prediction matrix*. Columns in this matrix are comments in $\mathcal{M}$, and rows are subreddit classifiers. Each cell [i,j] in the prediction matrix contains a *yes* or *no*, depending on what classifier clf_{S_i} predicted for comment m_j: *If it were hypothetically posted here, would it get moderated?*

not present in $\mathcal{M}$. Essentially, any comment that we crawl from Reddit, which is not removed by a moderator within 24 hours from the time of posting is added to our unmoderated comments corpus. These comments are collected similarly to the moderated comments, from the same set of subreddits, and through the same API. Using this data, we compile a dataset of unmoderated comments for all study subreddits.

4 METHOD: CLASSIFIERS FOR PREDICTING COMMENT REMOVALS

In this section, we detail the procedure used to train classifiers that can predict moderator removals within the study subreddits. Using the comments that were removed by moderators of each study subreddit, along with unmoderated comments collected from study subreddits, we train machine learning models to predict whether a comment posted on the subreddit will get moderated or not. An illustration of this approach is shown in Figure 3.

4.1 Building classifiers for study subreddits

Let us refer to each in-domain classifier built entirely using moderated and unmoderated comments from a single subreddit as a "subreddit classifier". We go on to build 100 such classifiers, one for each of our study subreddits. Each subreddit's classifier is trained on comments removed by moderators from the subreddit under consideration, along with an equal number of randomly sampled comments that were not moderated (at the time of our data collection).

Next, we describe the construction of our 100 subreddit classifiers, and evaluation of the in-domain classifiers through 10-fold cross-validation tests.

Table 1. Grid of parameter values used when running classification tests to find the best combination of parameter values for our models. The best values shown for all the parameters, found with a grid search, were used in all classifiers.

Parameter	Description	Range	Best value
lr	Learning rate	[0.05, 0.5]	0.05
epoch	Number of epochs	[25,30,50]	25
dim	Size of word vectors	[100,200]	200
ngram range	Max length of word ngrams	[1,2,3]	3
lowercase	Converting text to lowercase	[on,off]	on
punctuation removal	Remove punctuation in text	[on,off]	on
number removal	Remove numbers in text	[on,off]	on

4.1.1 Balancing datasets for each subreddit. We shuffle and balance each subreddit's dataset to ensure an equal number of comments from each class (moderated and unmoderated). Note that balancing the number of samples from each class likely does not mimic real-world situations. In general, moderated posts are less frequent than unmoderated posts. However, balancing across all conditions ensures that we can easily interpret model fits relative to one another.

4.1.2 `FastText` classifiers. `FastText` is a state-of-the-art library for text classification [3, 25]. It represents each instance by the average of vector representations for words and n-grams, which are short units of adjacent words. These "representation vectors" enable generalization to words and n-grams that are not encountered in the training data. Supervised training is used to estimate another set of vectors, per label, which characterize the classification rule. If learning is successful, then subreddits with similar moderation patterns will have similar classification vectors.

4.1.3 Parameter tuning using gridsearch. Using `FastText`, we build 100 in-domain subreddit classifiers, each trained on an equal number of moderated and unmoderated comments obtained from the study subreddit (i.e., binary classification with balanced classes). Like all classifiers, `FastText` has a number of parameters that must be tuned to achieve optimal performance. We tune these parameters by grid search, trying a large set of values, and selecting those which maximize the F1 (f-measure) across 10-fold cross-validation. The parameter space, along with the best performing parameter values are shown in Table 1.

4.1.4 Evaluation of in-domain subreddit classifiers. Using the best performing parameter values shown in Table 1, we train in-domain subreddit classifiers for each of the 100 study subreddits. In order to evaluate the subreddit classifiers, we perform 10-fold cross-validation tests using the balanced set of moderated and unmoderated comments collected from each study subreddit. The mean 10-fold cross-validation F1 score for the 100 study subreddits was 71.4%. This is comparable to the performance achieved in prior work on building purely in-domain classifiers to identify moderated comments within an online community [6, 9].

4.2 Compute agreement among subreddit classifiers' predictions

We obtain the predictions from each of the 100 subreddit classifiers for all moderated comments present in $\mathcal{M}$. The prediction from each subreddit classifier for a comment represents a probabilistic answer to the following question: *If this comment were posted on this subreddit, would the moderators remove it?*

Next, we compute the overall agreement among all subreddit classifiers' predictions for each comment present in $\mathcal{M}$. By overall agreement among subreddit classifiers for a comment, we refer to the number of classifiers that agree to remove the same comment. The output of this step is a

prediction matrix, with number of rows equal to the number of comments in $\mathcal{M}$ (2.8M), and the number of columns equal to the number of subreddits for which we have trained classifiers (100).

4.3 Methodological limitations

4.3.1 Access to only textual data. Our current method of data collection does not give us access to removed content in the form of pictures, GIFs or videos. As a result, we are not able to identify community expectations around multimedia content.

4.3.2 False negatives. Anecdotal evidence shows that not all comments posted on a subreddit have been seen by moderators [20]. This could lead to some *false negatives* (i.e., comments that should have been removed) being present in our collection of unmoderated comments, which could affect the classifiers we build. Future work can investigate this issue, and examine the amount of such comments that the moderators typically fail to see on Reddit.

4.3.3 Passive norms. Note that the classifiers learn about rules and norms that are actively enforced by moderators on a subreddit. It could be the case that there exist "passive norms" that have not needed to be actively enforced—no one has thought to violate those norms within the subreddit. For example, posting TV show spoilers may actually be considered a norm violation on many subreddits, but the classifiers may not identify that such a comment would be moderated on a specific subreddit if no one has posted such spoilers within that subreddit before. Such passive norms could serve as "blind spots" for the subreddit classifiers, and may be an intriguing avenue for future study.

5 METHOD: CLUSTERING SUBREDDITS AND EXTRACTING NORMS

Next, we identify subreddits where moderators enforced similar community norms, by finding clusters of subreddits that would moderate the same comments. An illustration of this approach is shown in Figure 4. Using the predictions obtained from the 100 subreddit classifiers for all moderated comments in $\mathcal{M}$, described in the previous section, we cluster the subreddits based on their agreement with respect to moderating comments. For each cluster of subreddits, we then extract the norms enforced by moderators of most of these subreddits. We employ open coding to qualitatively identify norm violations exhibited by comments predicted to be moderated by subreddit classifiers.

Note that an alternative scheme like matrix factorization (or topic modeling, or clustering) on the *comments themselves* would likely just group subreddits by content, rather than by moderation practices (i.e., the decision to moderate comments or not). We believe that this would be true even if we focused exclusively on moderated comments, since a norm-violating comment in, say r/nba, is still likely about basketball. There is one key aspect that the procedure described above would miss: the labeling associated with the moderated posts. The procedure we use in this work, on the other hand, focuses on moderated comments. We begin by identifying commonality in the language of moderated comments via the subreddit classifiers, and then use this commonality to cluster the subreddits, arriving at subreddits clustered by *moderation practices*. Finally, we return to the language of the moderated comments to analyze the topics that characterize the obtained clusters.

5.1 K-means clustering

We use the K-means clustering algorithm [21] to cluster subreddits based on their predictions on all removed comments present in $\mathcal{M}$. Here, we find coherent clusters of subreddits that would remove similar comments. Because the matrix of subreddit-comment predictions is large, we first reduce its dimensionality by performing Principal Component Analysis [24]. Intuitively, PCA reduces the

Table 2. Clusters obtained from K-means clustering, based on agreement among classifier predictions to remove comments. The subreddits in each cluster are ordered by cosine distance from their respective cluster's center. *Size* denotes the number of subreddits present in the cluster, *type* denotes the cluster type or type of "norm" that is shared by subreddits present in the cluster, and *name* denotes the assigned cluster number by which we will reference each cluster in further sections.

Cluster subreddits	**Name**	**Size**	**Type**
conspiracy, Android, atheism, Incels, PurplePillDebate, IAmA, canada, tifu, india, SubredditDrama, dataisbeautiful, pics, LifeProTips, hiphopheads, fantasyfootball, explainlikeimfive, worldnews, SandersForPresident	C_0	18	Meso
CanadaPolitics, spacex, changemyview, NeutralPolitics, personalfinance, AskHistorians, history, whatisthisthing, science, Games, philosophy, space, Futurology, syriancivilwar, legaladvice, PoliticalDiscussion, AskTrumpSupporters, TheSilphRoad, Christianity, DIY, OutOfTheLoop, UpliftingNews	C_1	22	Meso
DestinyTheGame, hearthstone, Overwatch, jailbreak, 2007scape, wow	C_2	6	Meso
CFB, me_irl, books, movies, nba, nfl, asoiaf, pokemon, MMA, relationships, AskWomen, food, pcmasterrace, Showerthoughts, GlobalOffensiveTrade, pokemongo, leagueoflegends, depression, gonewild, hillaryclinton, SuicideWatch, The_Donald, gaming, GlobalOffensive, anime, politics, photoshopbattles, television, ShitRedditSays, GetMotivated, aww, EnoughTrumpSpam, sex, gameofthrones, TwoXChromosomes, funny, nottheonion, europe, LateStageCapitalism, news, technology, soccerstreams, socialism	C_3	43	Meso
churning, NSFW_GIF, pokemontrades, nosleep	C_4	4	Meso
videos, OldSchoolCool, gifs	C_5	3	Meso
AskReddit	C_6	1	Micro
BlackPeopleTwitter	C_7	1	Micro
askscience	C_8	1	Micro
creepyPMs	C_9	1	Micro

size of the input matrix by iteratively computing a projection that explains the most variance in the input. We find that a projection on 81 dimensions is sufficient to explain 90% of the original variance. We then cluster the PCA-transformed predictions of the subreddit classifiers using K-means. We determine the number of clusters k by examining the mean silhouette coefficient [30]—the similarity of predictions within a cluster with respect to other clusters. We test the clustering algorithm with different initializations of K (ranging from 1 to 20), in order to identify the most stable configuration.

5.2 Clustering results

By increasing K from 2 to 20, we find that after an initial local maximum the coefficient peaks around $K = 10$, before degrading for higher values. Therefore, we cluster the predictions in $K = 10$

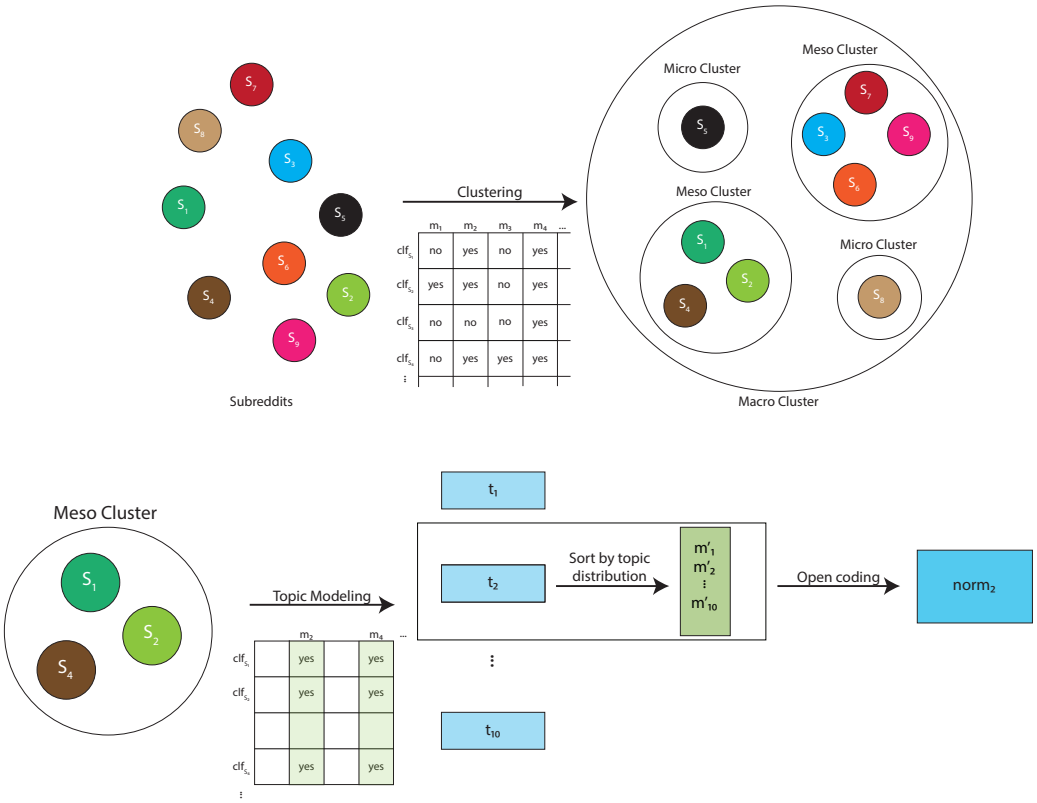

Fig. 4. Based on agreement among subreddit classifiers (e.g., clf_{S_k}) to remove comments (e.g., m_j), we cluster subreddits (e.g., S_k) into three different types of clusters: *macro*, *meso*, and *micro* clusters. For each cluster of subreddits, we perform topic modeling only on comments in $\mathcal{M}$ that the subreddit classifiers agreed to moderate, using the prediction matrix shown in Figure 3. Finally, we employ open coding to extract the norms violated by 10 comments that rank highly in the topics we identify. By repeating this procedure for the macro cluster containing all subreddits, and each cluster shown in Table 2, we extract *macro*, *meso*, and *micro* norms.

groups. The resulting 10 clusters are shown in Table 2, and the 2-D t-SNE [31] representation of the clusters is shown in Figure 5.

Based on the amount of agreement among subreddit classifiers, we identify three different levels of clusters among the study subreddits.

5.2.1 Macro cluster. First, we consider all 100 study subreddits to be part of *one large cluster*, so that we can identify comments that a large majority of subreddit classifiers belonging to this cluster agree to remove. These comments are highly likely to be removed by moderators of all study subreddits, when posted on their subreddit. Using the text contained in these comments, we will extract norms that extend across most study subreddits. We call these *macro norms*, as we observe them to be enforced by moderators of a large majority of our study subreddits.

5.2.2 Meso clusters. We identify six meso-clusters of subreddits (C_0 to C_5), obtained through K-means clustering, shown in Table 2. Moderators from all subreddits belonging to a cluster tend

to agree on what comments to remove from their subreddits (based on the predictions obtained from the subreddit classifiers). For each *meso cluster*, we identify comments that a large majority of subreddit classifiers belonging to this cluster agreed to remove, while subreddits that do not belong to the cluster agreed to not remove. For each comment, we compute the following ratio: fraction of subreddits within the cluster that agree to remove the comment (based on classifier predictions), normalized by the fraction of subreddits outside the cluster that agreed to remove the comment. Then, we rank all comments based on this computed ratio, and then pick only the top 1% out of all comments. These comments are highly likely to be removed only by moderators of subreddits present in the same cluster. Using the text in these comments, we will go on to qualitatively extract cluster-specific norms that extend across most study subreddits in the same meso cluster. We will call these *meso norms*, as we they are likely to be enforced by moderators of communities (subreddits) in the same meso cluster, but not on other parts of Reddit.

5.2.3 Micro clusters. Finally, we have the four micro-clusters (C_6 to C_9) obtained in Table 2, each containing a single, isolate study subreddit, in order to identify comments that are only removed by moderators of these individual subreddits. We identify comments that only the individual subreddit belonging to each micro cluster agreed to remove, while all other subreddits agreed to not remove. For each comment, we compute a similar ratio: fraction of subreddits within the cluster that agree to remove the comment (either 0 or 1 since there exists only one subreddit in the cluster), normalized by the fraction of subreddits outside the cluster that agreed to remove the comment. We rank all comments based on this computed ratio, and then pick only the top 1% comments. These are comments that violate highly specific norms that are enforced by moderators of micro cluster subreddits, while the same comments are not removed when posted on most other study subreddits. Using the text in these comments, we will qualitatively extract norms that are highly specific to each individual study subreddit. We call these *micro norms*, as we observe them to be enforced exclusively by moderators of individual subreddits.

Note: Subreddits are clustered based on the comments that their classifiers agree to moderate, and these are not necessarily representative of typical comments found on these subreddits. As a result, some of the obtained clusters may not be intuitive, and subreddits present in the same cluster need not appear to be topically similar. Instead, what we observe in these obtained clusters are subreddits that share similar moderation policies and norms. As mentioned in 5.1, the obtained clusters were determined to be the most stable configuration by examining the mean silhouette coefficient [30].

5.3 Norm extraction through topic modeling and open coding

As explained in the previous subsection, we identify clusters of subreddits that share norms among themselves at three different levels (macro, meso and micro).

5.3.1 Topic modeling. We next adopt a computational approach to reduce the dimensionality of our textual data. We employ topic modeling on the comments agreed to be moderated by subreddit classifiers belonging to each cluster to identify the underlying topics contained in these comments. We frame the task as follows:

Applying Latent Dirichlet Allocation (LDA) [2], we estimate topic distributions on the comments that have high agreement among classifiers belonging to the same cluster. We use LDA to estimate the topic distributions among 10 topics for each cluster. Every comment belonging to each cluster we analyzed is considered to be a document for this analysis. In further analysis, we tested by increasing the number of topics from 10 to 20 for LDA, but observe that no new types of norms emerged. As a result, we estimate topic distributions among 10 topics for each subreddit cluster.

5.3.2 Open coding for mapping topics to norm violations. Finally, we introduce a qualitative step, where we use open coding to manually code each topic by the norm violation it represents (in the form of a 1-2 line explanation behind the comment's removal). Using the topic distribution computed for all comments agreed to be removed by subreddits within each cluster, we identify 10 comments that ranked highly in each of the 10 topics obtained for the cluster. Then, three annotators independently code each topic by the norm violation it represents, using the 10 comments ranking highly in that topic as context. This way, we manually map all 10 topics (using 10 randomly sampled comments for context) to their respective norms for each cluster. Then, the three annotators come together to compare the norms they coded independently, and resolve any disagreements. By repeating this process for all clusters at the three different levels, we extract the macro, meso and micro norms contained in $\mathcal{M}$.

Through open coding, a total of 100 different topics were coded manually, and we observed the presence of 32 topics for which the annotators could not identify the exact norms being violated. This could arise from a number of different factors: computational noise introduced by the classifiers in the data; lack of background knowledge about the actual subreddits as outsiders; and, missing the context information for comments that were moderated. For example, some of these comments could have been removed by moderators due to reasons that are very highly context-specific to the type of discussions they were a part of. We discarded such topics for which we could not identify the exact norms being violated, and only present the norms that were identified and agreed upon by all three annotators[10].

5.4 Methodological limitations

Our findings hinge on the algorithms we use in our methods—the classifiers we train and the clustering algorithm we choose to employ can play a role in the types of norms we uncover. On the other hand, using these algorithms give us the ability to study site-wide norms holistically in a large-scale empirical manner, which is not possible to do by manual inspection alone.

5.4.1 Lack of context for removed comments. In our current analysis, we do not have access to the conversations surrounding the comments that were removed by the moderators of different subreddits. This lack of context for some of the removed comments could make interpreting the reasons behind moderator actions a hard task. Future work examining removed comments within the context of the larger discussions they are a part of could help understand moderator actions at a discourse-level. While it is true that we do not have context information for all moderated comments, the three independent annotators were able to agree on the norm violations represented by 68 out of the 100 topics that were coded.

5.4.2 Confounding factors. Note that we do not know the exact reason behind each moderator removal, and we do not account for differing levels of moderator activity within different subreddits. Currently, we do account for one common type of mass-removal: "children of the poisoned tree". Moderators sometimes remove all the children that were posted in response to comment that needs to be removed for violating community norms. The rationale behind this being, given that the parent needs to be removed, it is "safer" to remove its children, since there is high possibility of users responding in undesirable ways to an undesirable comment.

5.4.3 Treating auto-moderated and human-moderated comments equally. Our analysis treats auto-moderated and human-moderated comments equally when constructing norms for communities. We are currently unable to systematically determine whether comments were removed by AutoModerator or human moderators. In fact, the ways in which AutoModerator is used for

[10]The raw annotation data, with all three labels by independent coders, will be made available after blind review.

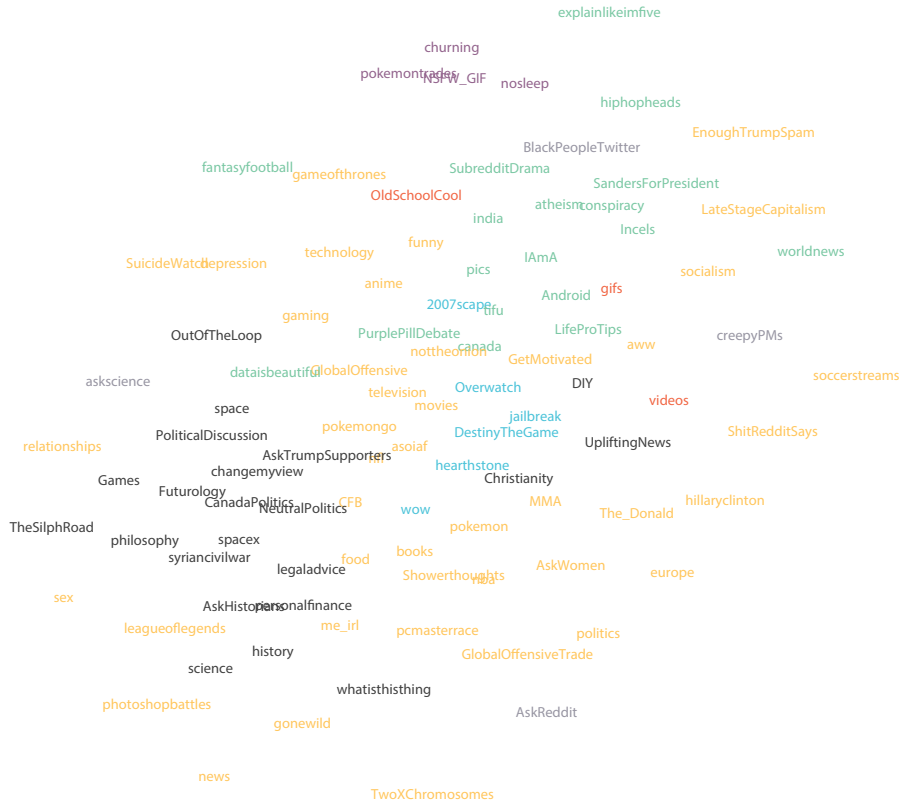

Fig. 5. 2-D t-SNE representation of the clusters, obtained from the high-dimensional space of subreddit classifier predictions. Intuitively, subreddits that are spatially nearby have similar moderation practices, according to the classifiers. Clusters are indicated by color, with all singleton subreddits shown in gray.

moderation varies from subreddit to subreddit. Anecdotally, we know that some subreddits remove all comments that trigger any of the hard-coded rules defined for the AutoModerator, while other subreddits use AutoModerator for triaging comments, which are subsequently reviewed by human moderators. Future work can examine the biases introduced by automated tools, like AutoModerator, that moderate based on hard-coded rules, when constructing community norms.

5.4.4 Temporal aspects of community norms. It is important to note that norms can change within and across communities over time, and tools that moderate automatically based on the "right" set of norms for a community must be flexible to change over time. Our current analysis presents a static snapshot of norms identified through moderator actions, and does not examine the temporal aspect of community norms. Future work may find traction exploring the temporal nature of community norms, examining how norms evolve within communities over time.

6 RESULTS

We identified 8 macro, 21 meso, and 15 micro norms by employing open-coding on comments agreed to be removed by subreddits from different clusters of subreddits. They are shown in Table 3, Table 4, and Table 5 respectively.

Table 3. Macro-norms extracted by analyzing comments that at least 96 out of 100 subreddit classifiers predicted to moderate from their respective subreddits. For each norm, we include an example comment found to be violating it.

Norm violations	Example comments
Using misogynistic slurs	*what a dumb cunt lol what a pussy*
Opposing political views around Donald Trump (depends on originating subreddit)	*stay classy trump supporters you bunch of worthless fucking pricks*
Hate speech that is racist or homophobic	*you're allow to swear on the internet you fucking [n-word]*
Verbal attacks on Reddit or specific subreddits	*drain the swamp, u/spez is a kek kekadooddledoo fuck reddit this site sucks*
Posting pornographic links	*you dont like senpai [URL]*
Personal attacks	*please kill yourself you useless sack of shit*
Abusing and criticizing moderators	*lets see if this gets deleted. fuck you r/news mods*
Claiming the other person is too sensitive	*fucking cry about it you fucking baby*

Table 4. Meso-norms extracted for Clusters C_0 to C_5 by analyzing comments agreed to be moderated by most subreddits within each cluster. For each norm, we include example comments and also the names of the clusters that enforce it.

Norm violations	Example comments	Clusters
Meme responses	*mitochondria is the powerhouse of the cell*	C_0
Comments that only express thanks	*thank you so much for sharing this*	C_0, C_5
Ad hominem attacks that demean and undermine users, based on flairs or usernames	*just looking at your rank flair I wouldn't really criticize*	C_0
Mocking the concept of safe space	*poor snowflake do you need a safe space*	C_0
Attempts to be funny, sarcastic, or make jokes	*its obvious god is really keen on what one eats and dinner etiquette*	C_1
Personal reactions, opinions	*and this is why i love science, always on the pursuit of knowledge*	C_1, C_4
Phatic talk	*they're making a new austin powers movie*	C_1
Outbound links to illegal live streams	*free live streaming chicago bulls los angeles lakers basketball*	C_2

Personal anecdotes (details about one's family, past events)	*According to my parents, my dad wanted to name me 'Taylor made' and my mom was like there is no way in hell you are doing that to my child. So Taylor with a normal first name was the compromise*	C_2, C_4
References to trading items	*i have a competitive shiny gengar which i could trade for that lugia if you're interested*	C_2
Expressing disagreement and criticisizing opinions shared by others	*i totally see where your coming from. i don't think however that we should approach this with a gradient...*	C_2, C_3
Talking about romantic relationships and sex	*just ask him out and see how it goes*	C_2, C_3
Mansplaining	*I'm not saying it was her goal. I'm saying her actions were akin to someone who had that goal*	C_2
Talking about guns	*laws vary by jurisdiction in a lot of places pointing the gun is automatically a threat but not pointing the gun is not automatically not a threat*	C_3
Excessive hedging	*maybe that is how you interpreted it but that is not necessarily what they meant*	C_3
Using Wikipedia articles and other web links to support arguments	*it was an acquired accent taught in schools source [Link to Wikipedia]*	C_3, C_4
Generalized complaining (e.g., electoral system, censorship, airport rates, etc.)	*yeesh airport rates are always silly but it s still disheartening to see a rate like that in any context*	C_3, C_4
Acknowledging a good point	*that's a valid point. honestly i had not thought about it that way before*	C_4
Links to promotional spam	*would you rate kenya coffee [Link to blog] ... as the best in the world or at least amongst the best*	C_5
Mocking religion and nationality	*but but but but but but but but islam is a religion of peace*	C_5
Hostility towards Muslims, and immigrants	*the country has to import rubbish from other countries. is that what we are calling it now*	C_5

Table 5. Micro-norms extracted for Clusters C_6 to C_9 by analyzing comments predicted to be moderated only by the individual subreddit within each cluster. For each norm, we include example comments and also the names of the clusters that enforce it.

Norm violations	**Example comments**	**Clusters**
Comments that only express thanks	(see above)	C_6, C_8
References to movies and TV shows	*on the other hand what other episode could it really be*	C_6
Offering commerce tips	*i could offer 25k so i think around somewhere there*	C_6, C_8

References to history	*perhaps the initials are emperor wilhelm, as in wilhelm ii who reinstitute it in 1914. that would also explain the crown*	C_6
Using Wikipedia links as source	(see above)	C_6
Personal reactions	(see above)	C_7
Guessing at other people's motives	*even by the fn their private views are likely different from their public views*	C_7
Talking about past regrets and lost opportunities	*wow i smoked pot for the first time at 13 and also dropped out of high school*	C_7
Merely indicating agreement conversation	*definitely i agree*	C_7
Personal anecdotes	(see above)	C_8
Diet advice, and pro-anorexic content	*i'm 153 lbs 5 9 and if i don't eat much a couple days in a row i can lose up to 5 or 6 lbs. once i've gone more than 3 or 4 days without much food. i completely lose my appetite and have to force feed. 1lb a day doesn't seem like much, i've lost up to 20 lbs in 3 weeks and that was when i was just eating when ever i was hungry*	C_8
High-school science theories	*when i was in highschool, i misunderstood the myth even more thinking that overnight a car battery would turn into some sort of acidic goo pile. i left a car battery on the front walk of my high school principal's house one night as a prank, again thinking he would come out the next morning to a pile of acidic slime. i wonder how confused he was to find a perfectly normal car battery in his yard the next morning*	C_8
Undermining and arguing against author opinions	*how is this disrespectful or hateful? would you remove my comment if it was criticizing the prevalence of homophobia related to christianity i don t think that's fair in the slightest*	C_9
Calling out previous authors for flaws	*i call bullshit on it*	C_9
Showing lack of confidence in one's own position	*i didn't say they were the same, i said a certain unnamed insult fits both*	C_9

6.1 Macro norms on Reddit

Working with moderated comments from 100 different communities on Reddit, we identified 8 macro norms that are enforced by the moderators on most subreddits.

Hate speech in the form of homophobic and racist slurs are considered as norm violations on most parts of Reddit. In addition, name-calling, use of misogynistic slurs, graphic verbal attacks, and distributing pornographic material are not condoned. Comments presenting opposing political views around Trump, either for or against depending on originating subreddit, are also removed by moderators. Such content could potentially lead to highly polarized comment threads, thereby hijacking ongoing discourse towards unrelated topics. This indicates that such comments are considered to be norm violations on Reddit because they hurt the process of discussion, and not necessarily because they are universally abhorrent. Another common norm violation is criticizing and abusing subreddit moderators, and most of the time, these are members of the community expressing their discontent with moderator actions (e.g., removing or promoting certain posts, lack of a formal escalation system, and the need for transparency in moderation). Sometimes, this discontent goes beyond certain specific subreddits, and users verbally attack Reddit (and its admins) due to a variety of reasons (e.g., policy change, banning communities, public statements, and so on). In such cases, moderators of most subreddits intervene and remove such comments.

6.2 Meso norms on Reddit

6.2.1 Cluster C_0. There are 18 subreddits present in this cluster, and they are on a range of topics (news, countries, politics, lifestyle, and so on). These communities have norms against ad hominem attacks, especially demeaning and undermining user opinion based on flairs or usernames. Moderators of these subreddits also remove comments mocking the concept of a safe space, and purely *meme* responses [37] (e.g., *"mitochondria is the powerhouse of the cell"*). Interestingly, comments that only express *thanks* are often observed to be removed by moderators. Though these comments serve a purpose for the two individuals that are part of the conversation, they do not necessarily add value, to other participants, within the context of the overall discussion. This type of removal could also be for archival reasons, where you want to minimize the amount of noise in current snapshots of the subreddit being archived for future references.

6.2.2 Cluster C_1. There are 22 subreddits present in this cluster, and most of them are subreddits that are known to be heavily moderated (e.g. r/NeutralPolitics, r/science, r/AskHistorians). Personal reactions, opinions, and (failed) attempts to make jokes or be sarcastic are considered to be violations of community norms on these subreddits. Additionally, references to movies, phatic talk, and comments that generally do not add value to ongoing conversation are removed by moderators.

6.2.3 Cluster C_2. There are 6 subreddits present in this cluster, and most of them are gaming-related subreddits (e.g., r/DestinyTheGame, r/Overwatch, r/wow). Moderators remove outbound links to (illegal) live streams and references to trading items (especially Pokemon). Other common removals include comments sharing personal anecdotes, stories about romantic relationships and sex. Mansplaining and criticizing opinions shared by other users are not condoned by these subreddits.

6.2.4 Cluster C_3. There are 43 subreddits present in this cluster, including highly popular subreddits focused on topics like politics (e.g., r/The_Donald, r/hillaryclinton), sports (e.g., r/NBA, r/nfl), and mental health (e.g., r/depression, r/SuicideWatch). Hedging language, criticizing other users' opinions, and the use of weblinks, including Wikipedia articles, to support arguments are not encouraged within these subreddits. Moderators also remove comments complaining about current state of things (e.g., electoral system, censorship, and so on).

6.2.5 Cluster C_4. There are 4 subreddits present in this cluster, and they are r/churning, r/NSFW_GIF, r/pokemontrades, and r/nosleep. Norm violations include complaining about the state of things,

and using Wikipedia articles to make a point. Moderators also remove comments that are personal reactions, personal anecdotes, or just acknowledging a good point.

6.2.6 Cluster C_5. There are 3 subreddits present in this cluster, and they are r/videos, r/OldSchoolCool, and r/gifs. Hostility towards muslims and immigrants, and mocking religion and nationality violate the norms of these communities. Moderators also remove links containing promotional spam, and low value comments that only express *thanks*.

6.3 Micro norms on Reddit

Micro norms are context-dependent, and highly specific to individual subreddits, and are not found to be widely enforced on most parts of Reddit. For instance, moderators of r/AskReddit, a Q&A forum, consider low value comments that express gratitude, contain movie or TV show references, and offering commerce tips as norm violations. In addition, references to historical events, and comments using Wikipedia links to support their arguments are removed by moderators, despite there being no written rules against them. r/BlackPeopleTwitter is intended for hilarious and insightful social media posts by black people, with an emphasis on hilarity.[11] As a result, posting personal reactions to issues, guessing at the motives of users, and talking about past regrets or missed opportunities are considered norm violations by the moderators. On a science Q&A forum promoting scientific literacy like r/askScience, posting personal anecdotes is against comment rules, as specified by subreddit moderators. We also observed that moderators do not tolerate high-school science theories, and diet advice (especially pro-anorexic content) in discussions. On a support subreddit like r/CreepyPMs, undermining, arguing against, and calling out flaws present in comments/post by previous authors are considered norm violations. Sometimes, comments showing a lack of confidence in one's own position are also removed by moderators.

7 DISCUSSION

Our findings describe the ecosystem of norms on Reddit. Some of the community norms we identified are mirrored in the written rules and guidelines provided by Reddit, or individual subreddits (an encouraging face validity sign); however, many are not. We also see many *unpublished* norms that are widely enforced by subreddit moderators.

7.1 Norms at different scales on Reddit

Our findings document the existence of norm violations that are universally removed by moderators of most subreddits. These include comments that contain personal attacks, misogyny, and hate speech in the form of racism and homophobia. The presence of these macro norms are in many ways encouraging, as they indicate that engaging in such behavior is considered a norm violation site-wide. We would argue that knowing about the presence of such site-wide norms could also help moderators of new and emerging communities shape their regulation policies during the community's formative stages, and feel more confident doing so.

We also documented norms that are local to specific groups of subreddits—the meso norms. For instance, sharing personal anecdotes, and posting links containing promotional spam are considered norm violations in certain clusters of subreddits (C_2, C_4, and C_5), while most other communities on Reddit do not consider such comments norm violations. We also found some meso norms that are seemingly counter-intuitive. For instance, comments expressing thanks, or acknowledging a good point are considered to be norm violations in clusters C_0, C_5, and C_4 respectively. Though these comments appear to be polite, and add value to one-to-one conversations between individual users, they may be perceived as *noise* or low-value comments by users trying to follow the larger

[11] https://www.reddit.com/r/BlackPeopleTwitter/

discussion. On the other hand, we observe that only certain clusters considered mansplaining (C_2), mocking religion and nationality (C_5), and hostility towards Muslims and immigrants (C_5) as norm violations. Despite being important societal issues, they do not appear to be norm violations on most parts of Reddit.

Finally, we observed the presence of highly specific micro norms that apply to individual subreddits. These are distinctive to the particular subreddits they emerge from, and are not widely enforced on most other parts of the site. For example, using Wikipedia as a source and presenting high-school science theories are considered to be norm violations within r/AskReddit, and r/askscience respectively, while most other parts of Reddit would not remove such comments. These idiosyncratic micro norms are important for understanding the reasoning behind moderator removals within individual communities—as well as understanding the range of norms on an umbrella site like Reddit.

7.2 Ethical considerations

We recognize that the use of "deleted data" (here in the form of moderated comments) is controversial territory in social computing research. We debated and discussed these issues with our local colleagues, remote colleagues, and our IRB before performing this research. In the end, we arrived at the conclusion that examining moderated comments provides invaluable insights about the governance of online communities, and as long as any downside risks are mitigated, those benefits outweighed the risks. For example, as we discuss next, we believe these findings may enable new mixed-initiative governance tools for online communities. We actively worked to minimize potential risks by not linking moderated comments back to their authors (who may not want to be immortalized in a research paper next to their norm violation). Moreover, we did not use *posts deleted by their authors* in this work, as those felt qualitatively different to everyone with whom we discussed this work. Finally, in an effort to protect Reddit itself from harm, we used only public data collected via Reddit's official API.

7.3 Theoretical implications

Norms play a key role in the governance of online communities [29]. Norms can be nested, in that they can be adopted from the general social context (e.g., use of pejorative adjectives are rude), or from Reddiquette[12], and more general internet comment etiquette (e.g., using all caps is equivalent to shouting at someone). Yet, norms for what is considered to be acceptable can vary significantly from one community to another, thereby making them challenging to study at scale. Through our work, we presented an empirical description of an ecosystem of community norms on Reddit, and our findings shed light on *what Reddit values*, and how widely-held these values are. We believe this is the *first large-scale study* of norms across disparate online communities.

Despite having established moderation strategies, including rules and guidelines, in place to regulate subreddits, bad behaviors continue to remain a challenge for online communities [17, 39, 41]. In the context of Reddit, rules and norms are interrelated. Moderators create formalized rules and guidelines for the front-stage of their subreddits, based on the norms they enforce in the back-stage. In our work, we identified norms as the emergent themes contained in the record of moderated comments. We observed that some of the norms we identified may overlap with outward-facing subreddit rules, but a far greater proportion of them do not. Future work could examine this apparent divide between the formal rules and informal norms enforced by moderators in online communities in greater detail. An understanding of the ecosystem of norms within online communities known

[12] https://www.reddit.com/wiki/reddiquette

to be successful in regulating behaviors could provide an empirical understanding of the driving factors behind effective online governance.

7.4 Design implications

7.4.1 Implications for online communities. For the design of online communities, it may be possible to use the frame of macro, meso, and micro norms to derive normative guidelines for new and emerging online communities. That is, the norms identified in this work may serve as sensible defaults for a new online community. On the other hand, some norms are problematic (e.g., *do not criticize mods*, *do not express thanks*) and suggest challenges for designing large-scale discussion systems. *How do you support dyadic relational maintenance without interfering in the larger discussion? How do you provide a place for discussion and arbitration of mod actions?*

For established online communities, an understanding of the macro, meso and micro norms on Reddit could help moderators reflect on the norms they typically enforce within their subreddits. Moderators can adopt existing norms from other communities known to be successful in regulating behaviors (e.g. r/AskHistorians, r/askscience, and r/NeutralPolitics). This could also help train new moderators by surfacing the implicit norms in the community. For new communities, we believe these macro norms (and some meso norms, depending on the community) may serve as sensible defaults for regulating behavior.

7.4.2 Designing automated moderation tools. When designing automated moderation tools for online communities, it is important to take the community's norms into account. Moderators play a key role in governing online communities, and some of them have been doing their jobs for an extended period of time. By examining what the moderators actually remove, we can build better tools for triaging content that violates the community's norms. As a first step, we examined what this space of online norms looks like empirically, by analyzing actual comments removed by moderators on Reddit. We observed that not all of the comments that get moderated are abusive or hateful in nature. There exist many other non-trivial, community-specific norms that get violated, resulting in moderator removals. Given that different communities care about different sets of norm violations, the severity of infractions can also be significantly different given the context or the nature of the topic of discussion (e.g., sensitive topics around politics or mental health would require the moderators to be on less tolerant of trolling or vitriol). These nuances are important to take into account, as platforms and researchers are doubling down on machine learning-based approaches toward moderation.

7.4.3 Classifiers that learn from other communities' norms . Finally, the discovery of widely overlapping norms suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values. We observed that the F1 scores obtained for relatively smaller subreddits (with less than 5000 removed comments) was approximately 68%, despite using state-of-the-art classifiers. This indicates the potential for using cross-community data to augment and improve completely in-domain classifiers. By understanding the types of norms that are valued by the target community, researchers could use classifiers trained on other source communities that share similar community norms or values.

One way to operationalize this idea is take into account the general agreement between the source and target community classifiers, measured by the number of comments both classifiers agree to moderate. By weighting out-of-domain classifier predictions with this inter-subreddit agreement measure, we could perhaps build cross-community classification frameworks for automated moderation that employ the scaffolding of other communities' norms.

8 CONCLUSION

We examined community norms on Reddit in a large scale, empirical manner. Using computational and qualitative methods, we examined over 2 million comments removed over 10 months by moderators of 100 top subreddits. We identified three types of norms within Reddit: *macro* norms which are universal to most parts of Reddit; *meso* norms which are shared across certain groups of subreddits; and *micro* norms which are highly specific to individual subreddits. We argued that our findings represent the first large-scale study of norms across disparate online communities, given the size and diversity of Reddit's user base. We concluded by reflecting on the apparent sharing of norms among distinct online communities, discussing implications for theory, and the design of internet communities more broadly.

9 ACKNOWLEDGMENTS

We thank the social computing group at Georgia Tech, and the Social Media Research Lab (SMRL) at University of Michigan for their valuable inputs that improved this work. Chandrasekharan and Gilbert were supported by the National Science Foundation under grant IIS-1553376. Eisenstein was supported by the National Science Foundation under award RI-1452443, and by the National Institutes of Health under award number R01GM112697-01.

REFERENCES

[1] Michael S Bernstein, Andrés Monroy-hernández, Drew Harry, Paul André Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *In Proc. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).

[4] Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. ACM, 1–10.

[5] Catherine Buni and Soraya Chemaly. 2016. The Secret Rules of the Internet. *The Verge* (2016). http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

[6] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.

[7] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1201–1213.

[8] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[9] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM.

[10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Web and Social Media*.

[11] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Ninth International AAAI Conference on Web and Social Media*.

[12] Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*. Vol. 24. Elsevier, 201–234.

[13] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* 58, 6 (1990), 1015.

[14] Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. (1998).

[15] Crowdsourcing Civility. 2014. A Natural Experiment Examining the Effects of Distributed Moderation in Online Forums/C. Lampe, P. Zube, J. Lee et al. *Government Information Quarterly* 31, 2 (2014), 317–326.
[16] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying.. In *The Social Mobile Web*. 11–17.
[17] Bruce Drake. 2014. The darkest side of online harassment: Menacing behavior. *Pew Research Center, http://www.pewresearch.org/fact-tank/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/* (2014).
[18] Maeve Duggan. 2014. Online harassment: Summary of findings. *Pew Research Center, http://www. pewinternet. org/2014/10/22/online-harassment/(accessed 02 June 2015)* (2014).
[19] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the 2018 AAAI International Conference on Web and Social Media*. ICWSM.
[20] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 803–808.
[21] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.
[22] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). http://firstmonday.org/ojs/index.php/fm/article/view/8232
[23] 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 (2018), 12.
[24] Ian T Jolliffe. 1986. Principal component analysis and factor analysis. In *Principal component analysis*. Springer, 115–128.
[25] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 427–431.
[26] Ruogu Kang, Laura Dabbish, and Katherine Sutton. 2016. Strangers on Your Phone: Why People Use Anonymous Communication Applications. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 359–370.
[27] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012).
[28] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 543–550.
[29] Lawrence Lessig. 1999. *Code and other laws of cyberspace*. Vol. 3. Basic books New York.
[30] R Lletı, M Cruz Ortiz, Luis A Sarabia, and M Sagrario Sánchez. 2004. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta* 515, 1 (2004), 87–100.
[31] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[32] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2857–2866.
[33] Elinor Ostrom. 2015. *Governing the commons*. Cambridge university press.
[34] Jessica Annette Pater, Yacin Nadji, Elizabeth D Mynatt, and Amy S Bruckman. 2014. Just awful enough: the functional dysfunction of the something awful forums. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2407–2410.
[35] Ari Schlesinger, Eshwar Chandrasekharan, Christina A Masden, Amy S Bruckman, W Keith Edwards, and Rebecca E Grinter. 2017. Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 6912–6924.
[36] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 111–125.
[37] Limor Shifman. 2014. *Memes in digital culture*. Mit Press.
[38] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.
[39] Nitasha Tiku and Casey Newton. February 4, 2015. Twitter CEO:"We suck at dealing with abuse.". *The Verge* (February 4, 2015). http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the
[40] Harry Charalambos Triandis. 1994. Culture and social behavior. (1994).

[41] Adam Mosseri (VP, News Feed). December 15, 2016. Addressing Hoaxes and Fake News. *Facebook Newsroom* (December 15, 2016). https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

[42] Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An Examination of Regret in Bullying Tweets.. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 697–702.

Received April 2018; revised July 2018; accepted September 2018