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

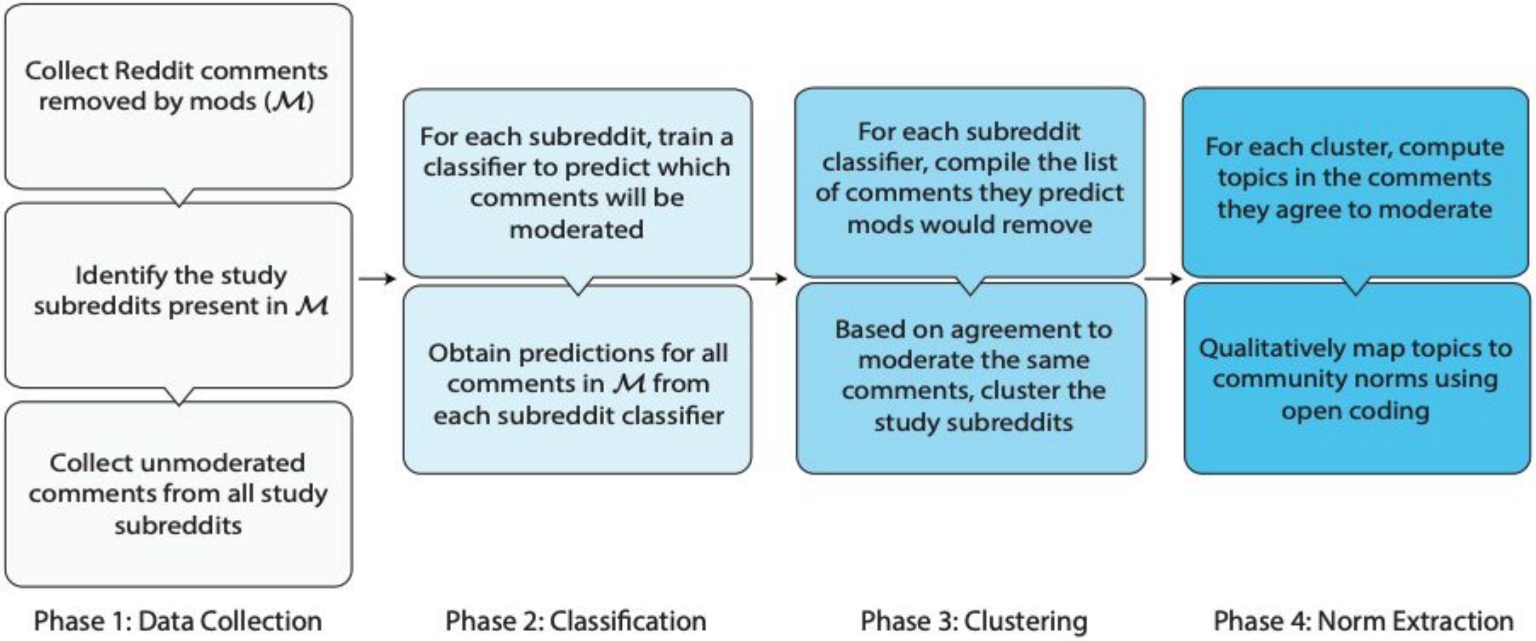
ESHWAR CHANDRASEKHARAN, MATTIA SAMORY, SHAGUN JHAVER, HUNTER CHARVAT, AMY BRUCKMAN, CLIFF LAMPE, JACOB EISENSTEIN, ERIC GILBERT

Presenter: Sifat Redwan

Motivation

- Norms are unwritten social expectations that guide behavior in online communities.
- **Norms** are different then **rules** as rules tend to be explicit, norms are emergent, arise from interaction over time, and respond to current demands on a community.
- Norms vary widely across communities in Reddit.
- Norms play an important role in guiding acceptable behaviors, and is key to governing online communities.
- This paper presents the first large-scale study of community norms across communities and shed light on what Reddit values, and how widely-held those values are.

Flowchart of Research Pipeline



Collect Reddit comments removed by mods ($\mathcal{M}$)

Identify the study subreddits present in $\mathcal{M}$

Collect unmoderated comments from all study subreddits

Phase 1: Data Collection

For each subreddit, train a classifier to predict which comments will be moderated

Obtain predictions for all comments in $\mathcal{M}$ from each subreddit classifier

Phase 2: Classification

For each subreddit classifier, compile the list of comments they predict mods would remove

Based on agreement to moderate the same comments, cluster the study subreddits

Phase 3: Clustering

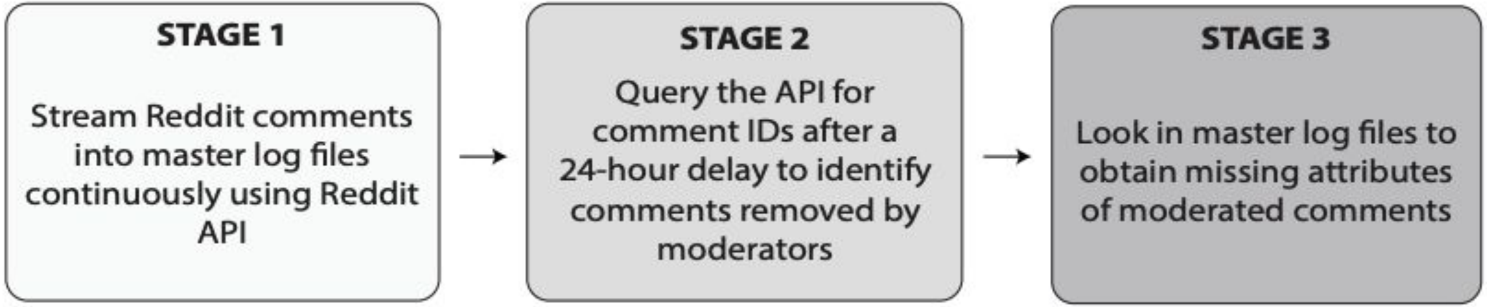
For each cluster, compute topics in the comments they agree to moderate

Qualitatively map topics to community norms using open coding

Phase 4: Norm Extraction

M denotes all moderated Reddit comments

Data collection



STAGE 1

Stream Reddit comments into master log files continuously using Reddit API

→

STAGE 2

Query the API for comment IDs after a 24-hour delay to identify comments removed by moderators

→

STAGE 3

Look in master log files to obtain missing attributes of moderated comments

Stage 1: Reddit comments were streamed into a master log file using the Reddit streaming API on a continuous basis.
Stage 2: After a 24-hour delay the removed comments were re-queried from the master log file. Comments that are removed are replaced with [“removed”].
Stage 3: A final look-up was performed for the moderated comments to obtain missing field like ‘body’, ‘subreddit’, ‘author’.

- 4,605,947 moderated comments were collected in the comment corpus (M).
- Master log file are a centralized record that continuously logs and stores data.

Preprocessing Moderated Comments in M

- **Automoderator replies:** All comments by automoderator were removed.

- **Discarding replies to moderated comments:** Referred as ‘children of the poisoned tree’. These replies were also removed from M. 1,051,623 comments were removed by this process.

- **Study Subreddits:** All subreddits with fewer than 5,000 moderated comments were removed.

- **Non- English subreddits:** All comments from any non-English subreddits were removed.

Finally, 2.8M moderated comments remain from 100 subreddits. These are called ‘**study subreddits**’ in the paper.

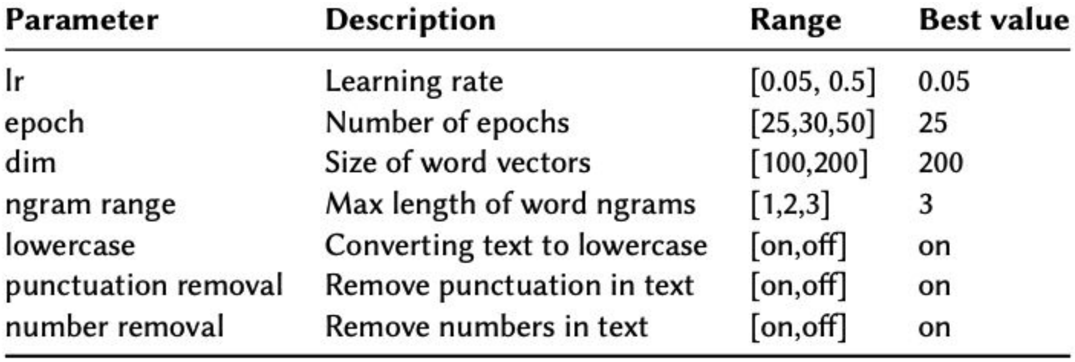
Building unmoderated comment dataset

- Any comment that **remains online** after 24 hours is considered **unmoderated**. These comments are separated from the removed ones and stored in an unmoderated comments dataset.

- The dataset of unmoderated comments includes all such comments from the same subreddits, collected using the same API and method.

Classifiers for predicting comment removals

- Machine Learning (ML) models were trained using the moderated and unmoderated comment for each study subreddit, to predict whether a comment posted on the subreddit will get moderated or not.
- For each subreddit, a classifier is built. So, 100 subreddit classifiers.
- **FastText classifiers** were used
- Parameter tuning using gridsearch
- Evaluation with 10-fold CV.
- Achieved an F-1 score of 71.4%.



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

Computing agreement among classifiers and Clustering

- Predictions from each of the 100 subreddit classifiers were obtained for all moderated comments present in M.

If this comment were posted on this subreddit, would the moderators remove it?

- Overall agreement was computed among the classifiers predictions for each comment present in M. That means the number of classifiers that agree to remove the same comment.
- Based on the agreement of the classifiers subreddits were clustered.
- For the clusters norms were extracted using open coding and qualitative method.

Clustering results

K-means clustering were employed with k=10. Three clusters were identified:

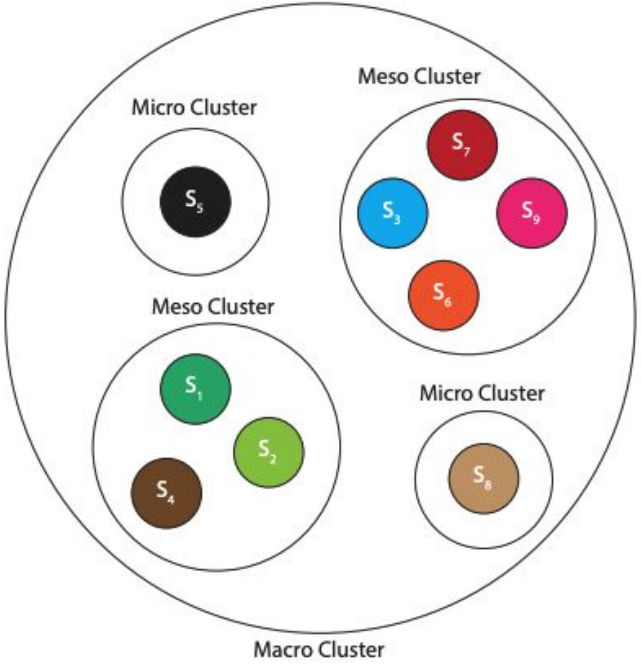
Macro Cluster: All 100 study subreddits are part of this cluster.

Meso Cluster: Subreddits that enforce similar norms within a specific (r/science, r/OverWatch, r/depression).

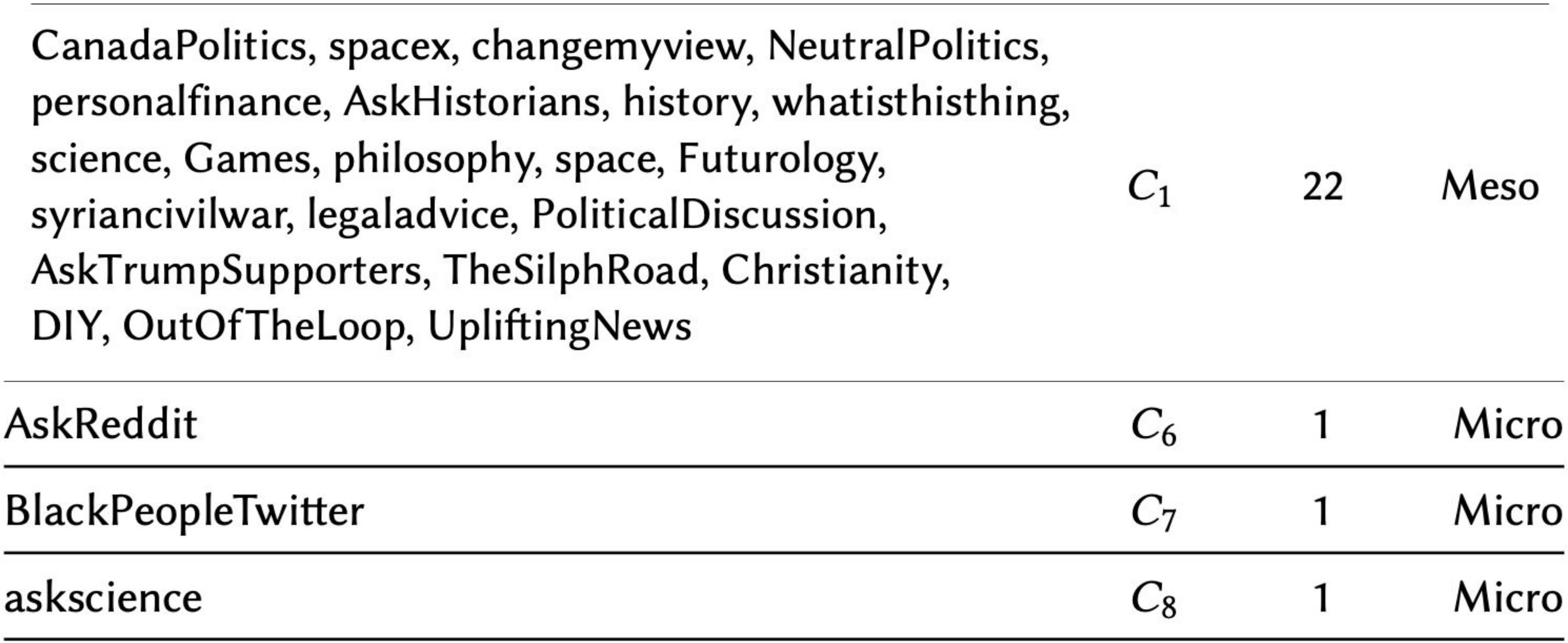
Micro Cluster: Highly specific norms unique to individual subreddits,

(r/askScience, r/BlackPeopleTwitter, r/CreepyPMs)

Moderators from all subreddits belonging to a cluster tend to agree on their subreddits.



Results after Clustering



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
AskReddit	C_6	1	Micro
BlackPeopleTwitter	C_7	1	Micro
askscience	C_8	1	Micro

Norm extraction through topic modeling and open coding

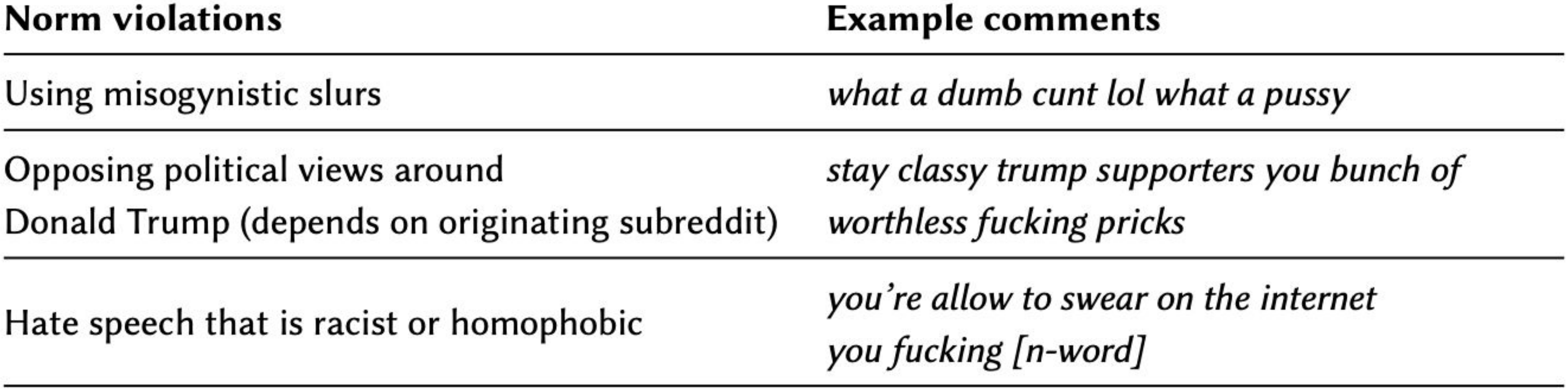
Topic Modeling: Latent Dirichlet Allocation (LDA) was applied to estimate the topic distributions on the comments to subreddits in the same cluster and top **10** topics were identified for each cluster.

Open Coding: This qualitative step involves manually coding each topic by the norm violation it represents. First, 10 highly ranked comments were sampled for the **10** topic. So, 100 topic coded. Then, three annotators manually coded norm violations to compare the norms and resolve any conflicts. 32 topics were discarded due to unclear norm violations.

LDA is a probabilistic topic modeling algorithm used to uncover hidden topics in a collection of text documents.

Results

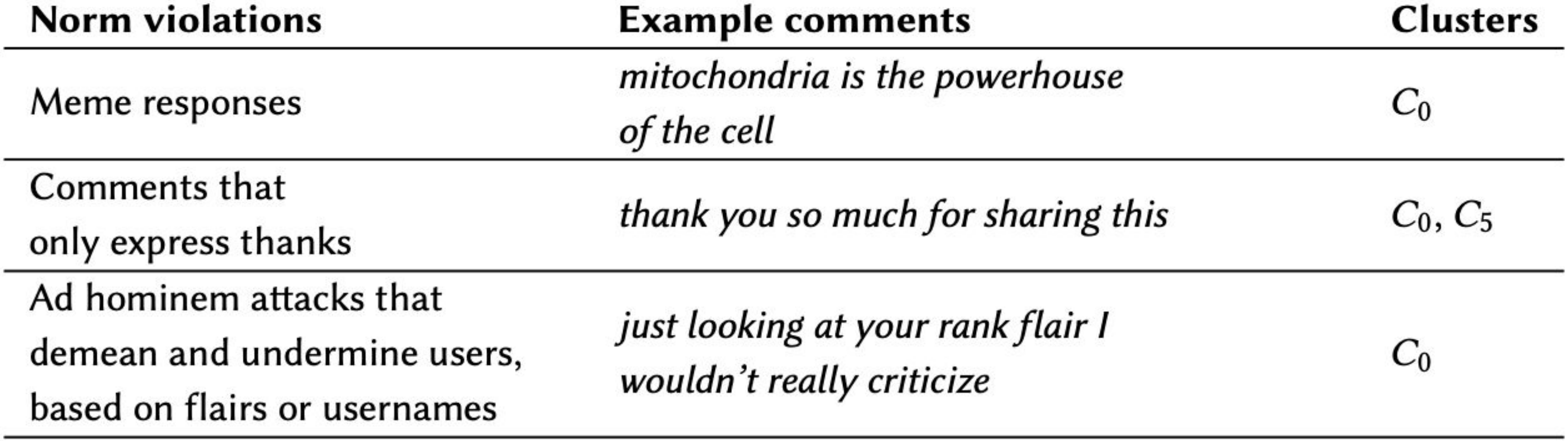
Macro Norm Violations: Hate speech, racist and misogynistic slurs, graphic verbal attacks, criticizing or abusing moderators, verbally attacking Reddit admins



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]*

Results

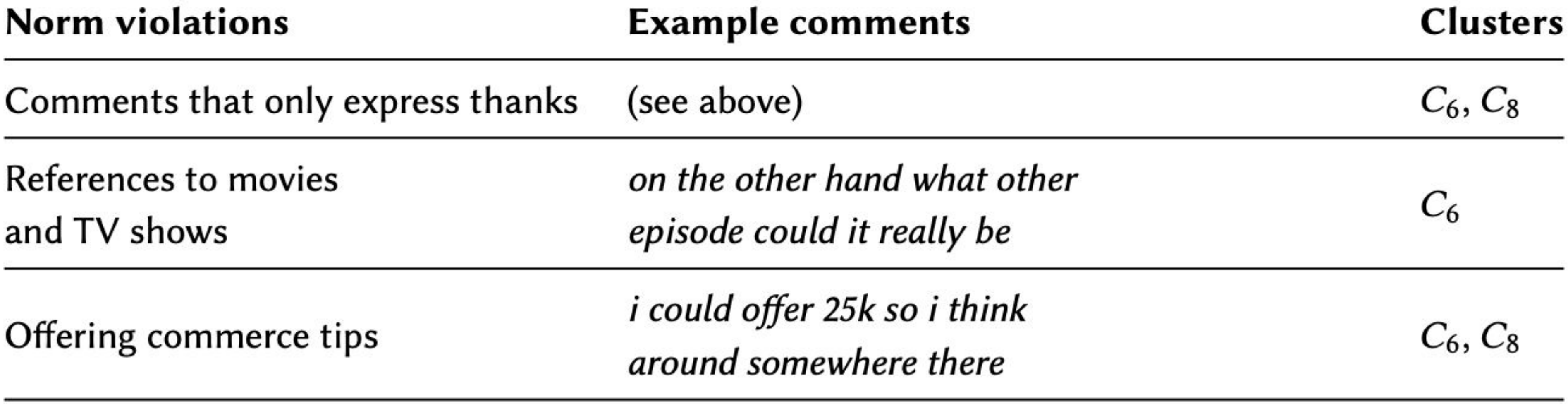
Meso Norms: Ad hominem attacks, sometimes *‘thanks’* are moved, purely meme response. *"I don’t know how to do taxes, but at least I know mitochondria is the powerhouse of the cell."*



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

Results

Micro Norms: Context dependent and highly specific to subreddits. **r/ AskReddit:** Low values comments like expressing gratitude, movie or TV show references, comments using wikipedia links. **r/askScience:** personal anecdotes, **r/CreepyPMs:** undermining, arguing



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

Limitations

- Lack of context for removed comments
- Confounding factors
- Temporal Aspects of community norms
- Access to only textual data
- Passive norms
- Despite using state of the art classifier, F-1 score was 71.4%.

Contributions

Implications for online communities

Designing automated moderation tools

Classifiers that learn from other communities norms

Summary

- Norms are key to how online communities are governed. They emerge from interactions between people and vary across community.
- This paper studies community norms on Reddit in a large-scale
- 2.8M comments removed by moderators were studied of top 100 subreddits.
- Three types of norms were found from these comments
- These findings suggest what Reddit values as norms
- This paper presents the first large-scale study of norms in Reddit.

Questions