Supplement for Measuring User-Moderator Alignment on r/ChangeMyView

January 16, 2023

1 Summary

In this supplement we provide a few additional materials to support our paper. First, we provide a detailed list of variables collected during our initial mod-log and mod-queue scrape. Second, we discuss in more detail the results of the moderator survey. Then, we describe how write-in responses were parsed for the practice-awareness and practice-support tasks. Next, we provide the results of the data simulations that motivated our sample size choice. Finally, we provide a table of coefficients for each MRP regression conducted, as well as user rating co-occurrence matrices for the practice-awareness and rule application tasks.

2 Moderation Action Dataset

Table 1: List of variables recorded for each `CMV` comment posted during the 3-month data collection period.

Variable Name	Source	Variable Description
`comment_id`	Comment stream	The ID associated with the comment in the Reddit API. To preserve user privacy, this field was dropped from the dataset after all relevant analysis was conducted
`comment_body`	Comment stream	The text of the comment
`post_title`	Comment stream	The title of the post associated with the comment
`post_body`	Comment stream	The main text of the post associated with the comment
`user_reports`	Mod-queue	A list of user reports issued against the comment. For comments collected prior to August 10, this list contains pairs of report reasons, and a count of their occurrences. For comments collected after August 10, this list contains pairs of report reasons, and the times at which they were recorded.
`mod_reports`	Mod-queue	A list of moderator reports issued against the comment. Each entry contains a report reason, a time of recording, and a salted cryptographic hash of the reporting moderator's username.
`mod_actions`	Mod-log	A list of moderator actions taken against a comment. Each entry contains an action type (typically "remove" or "approve"), a time of recording, and a salted cryptographic hash of the acting moderator's username.
`is_op`	Comment stream	A binary variable indicating whether the author of the comment matches the author of the associated post.
`is_reply`	Comment stream	A binary variable indicating whether the comment was a reply to another comment.
`is_reply`	Comment stream	A binary variable indicating whether the comment was a reply to another comment.
`flair`	Comment stream	A variable containing the number of deltas the comment author had accrued at the time of commenting
`time`	Comment stream	The time the comment was posted
`author_name`	Comment stream	The time between the creation of the associated post, and the creation of the comment

3 Moderator Survey Results Description

The results of the moderator were used to determine whether adequate context was provided for each comment. We specifically wanted to identify instances where moderators themselves needed additional information to make a decision, since if that were the case, users would likely need additional information as well.

Across the 268 ratings supplied in the moderator survey, raters said they might seek additional context before making a decision in 35/268 cases (13%). In only 2/268 of these cases did respondents say they were unable to make a decision at all without additional context. The most frequent additional context requested was additional comments associated with the post (e.g. comments higher up or further down in the thread, N=23). Other requests were whether or not the commenter was the OP (N=4), the commenter's post history (N=3), and information about why a comment was reported, if applicable (N=3).

In 12/268 cases, moderators said they might consult with others before making a final decision. In 3/268 cases, moderators said they would ask other moderators whether removing the comment was appropriate. In 4/268 cases, moderators said they would remove the comment, but consult others about whether a ban was appropriate. In 3/268 cases, moderators said they would leave the decision up to someone who felt more strongly about the right course of action. In 2/268 cases, moderators said they would make a record of the decision for other mods to look at, but would still go through with the action.

Given the relative infrequency of these cases, we decided that the individual survey format, though limited, provided enough context to compare user and moderator opinion. Although being able to provide additional comments from a thread for context would have been ideal, introducing this extra information would create additional complexity for survey participants.

While the results of the previous subsection did not lead to any changes to the user survey content, the user survey interface was redesigned. This was because two moderators had expressed some initial confusion regarding the layout of information in the survey interface. Since we had an open channel of communication with moderators, we were able to resolve these confusions quickly when the moderators were filling out the survey. However such communication would not be feasible during the deployment of the survey, motivating our redesign.

4 Write-in Response Parsing

In most cases, participants expressed an alternate action that either explicitly included removal (e.g. "the comment should be removed and the user should be banned") or explicitly did not include removal (e.g. "the user should just be downvoted"). In these cases responses were treated as equivalent to "The comment should be removed" or "No action should be taken" respectively. In other cases, users expressed a preference for removal, but only conditional on additional context (e.g. "If the user is the OP, this comment should be removed"). These responses were treated as "Unsure/Need more context." Occasionally, responses indicated a preference for removal, but only conditional on additional actions being taken by the moderator (e.g. "the mods should issue a warning, and then consider removal if the behavior persists") – in these cases the response was considered equivalent to "No action is necessary", since they do not express an immediate preference for removal. In all other cases the response was considered equivalent to "No action is necessary."

After remapping was conducted, all responses either corresponded to an express preference for immediate removal, uncertainty at the right course of action, or an express preference against immediate

removal (though potentially a preference for removal conditional on additional actions being taken).

5 Simulation Results

To determine a reasonable sample size, we conducted simulations to assess how precisely our model could infer practice-support. Recall that our primary practice support measure is the correlation across comments between the proportion of users who support removal and the actual binary moderator decision made. In each simulation, we first select a number of comments to get rated per rule n and a number of ratings to get for each comment k. Because our model is generative, we can simulate data according to the model with a fixed set of parameters, and then check how precisely our inference procedure recovers those parameters. For each combination of n and k we simulate 5 rules, 5 times, according to one of the following regimes:

1. **Low correlation regime:** Data is generated such that the latent variables underlying the user and mod labels have a correlation of .3
2. **Medium correlation regime:** Data is generated such that the latent variables underlying the user and mod labels have a correlation of .5
3. **Perfect Alignment regime:** Data generated under this regime does not follow our model. In this case, moderator labels always follow the majority opinion expressed by users

In each simulation, 3 rules followed the medium correlation regime, 1 rule followed the low correlation regime and 1 rule followed the perfect alignment regime. We then computed the coverage and width of the credible intervals for the practice support measure inferred by the model. Because the "perfect alignment" regime is not actually generated by our model, our inference of the practice-support measure for this rule was usually biased downward (hence why our coverage sits just below 80%). We quantify this bias by also computing the distance between our inferred practice support credible interval and the ground truth value in each simulation.

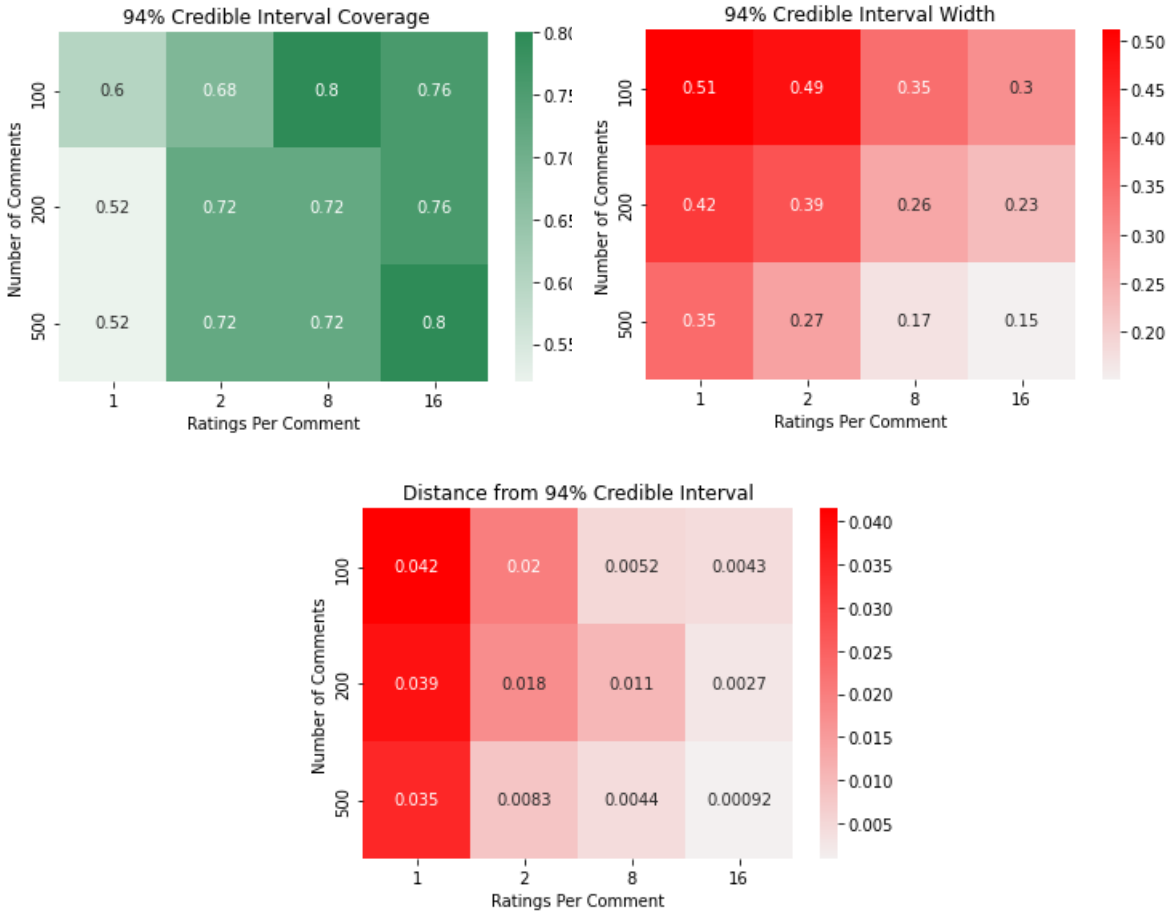

Figure 1: The coverage, width, and average bias of our model's practice-support credible intervals across simulations. One can observe that the inference is very low quality with a single rating per comment. With respect to decreasing interval width, the benefits of getting more comments tended to outweigh those of getting more ratings per comment, without sacrificing much in terms of coverage.

6 MRP Coefficients and Additional Co-Occurrence Matrices

Table 2: Regression coefficients for the MRP model in the policy awareness task. Each cell contains a 95% CI for the estimated parameter.

Variable Name/Model	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
# Comments (General)	(-0.04, 0.03)	(-0.02, 0.03)	(-0.00, 0.05)	(-0.06, 0.00)	(-0.016, 0.04)
# Comments (CMV)	(0.36, 1.20)	(-0.14, 0.34)	(0.21, 0.69)	(0.07, 0.96)	(-0.04, 0.47)
# Removals (General)	(-0.46, 0.20)	(-0.16, 0.36)	(-0.18, 0.29)	(-0.38, 0.42)	(-0.09, 0.45)
# Removals (CMV)	(-0.29, 0.63)	(-0.03, 0.79)	(-0.30, 0.18)	(-0.50, -0.26)	(-0.45, 0.04)
Account Age	(0.03, 0.53)	(-0.23, 0.24)	(-0.08, 0.30)	(-0.21, 0.40)	(0.12, 0.54)
Moderator Status Intercept[0]	(-0.95, 2.26)	(-1.10, 2.09)	(-1.47, 1.74)	(-0.60, 2.79)	(-0.94, 2.26)
Moderator Status Intercept [1]	(-0.67, 2.52)	(-1.12, 2.09)	(-1.10, 2.19)	(-0.78, 2.65)	(-1.20, 2.00)

Variable Name/Model	Fake Rule 1	Fake Rule 2	Fake Rule 3	Fake Rule 4	Fake Rule 5
# Comments (General)	(-0.49, -0.05)	(-0.72, -0.31)	(-0.60, -.06)	(-0.42, -0.00)	(-0.67, -0.23)
# Comments (CMV)	(-0.581, -0.157)	(-0.850, -0.458)	(-0.567, -0.137)	(-0.429, 0.000)	(-0.710, -0.281)
# Removals (General)	(-0.14, 0.30)	(-0.22, 0.18)	(0.03, 0.50)	(-0.20, 0.24)	(-0.07, 0.33)
# Removals (CMV)	(-0.04, 0.39)	(-0.06, 0.34)	(-0.49, 0.08)	(-0.31, 0.19)	(-0.06, 0.39)
Account Age	(-0.25, 0.13)	(-0.29, 0.08)	(-0.36, 0.05)	(-0.26, 0.11)	(-0.26, 0.13)
Moderator Status Intercept[0]	(-1.83, 1.38)	(-1.66, 1.51)	(-1.95, 1.28)	(-1.84, 1.36)	(-1.77, 1.34)
Moderator Status Intercept [1]	(-1.86, 1.40)	(-1.65, 1.53)	(-2.23, 1.07)	(-2.06, 1.13)	(-1.74, 1.42)

Table 3: Regression coefficients for the MRP model in the policy support task. Each cell contains a 95% CI for the estimated parameter.

Variable Name/Model	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
# Comments (General)	(-0.06, 0.18)	(-0.10, 0.11)	(-0.09, 0.12)	(-0.07, 0.17)	(-0.14, 0.09)
# Comments (CMV)	(0.08. 0.28)	(0.07, 0.26)	(0.09, 0.26)	(0.06, 0.26)	(0.05, 0.24)
# Removals (General)	(-0.29, -0.56)	(-0.31, -0.09)	(-0.29, -0.08)	(-0.31, -0.08)	(-0.31, -0.09)
# Removals (CMV)	(-0.26, -0.-5)	(-0.35, -0.12)	(-0.27, -0.-7)	(-0.24, -0.01)	-0.25, -0.06)
Account Age	(-0.03, 0.01)	(-0.03, 0.01)	(-0.03, 0.01)	(-0.03, 0.01)	(-0.03, 0.00)
Moderator Status Intercept[0]	(-1.07, 2.21)	(-1.42, 1.61)	(-2.01, 1.15)	(-0.86, 2.37)	(-1.28, 1.71)
Moderator Status Intercept [1]	(-0.25, 3.11)	(-0.99, 2.021)	(-1.82, 1.36)	(-0.82, 2.46)	(-1.01, 1.97)

Table 4: Regression coefficients for the MRP model in the practice awareness task. Each cell contains a 95% CI for the estimated parameter.

Variable Name/Model	Mod Survey Comments	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
# Comments (General)	(-0.62, 0.77)	(-0.17, 0.09)	(-0.12, 0.14)	(-0.10, 0.16)	(-0.28, 0.10)	(-0.17, 0.10)
# Comments (CMV)	(-0.19, 0.99)	(-0.10, 0.23)	(0.04, 0.38)	(-0.12, 0.19)	(-0.15, 0.34)	(-0.04, 0.36)
# Removals (General)	(-1.86, -0.27)	(-0.12, 0.15)	(-0.26, 0.02)	(-0.06, 0.21)	(-0.14, 0.31)	(-0.26, 0.05)
# Removals (CMV)	(0.35, 1.51)	(-0.14, 0.01)	(-0.10, 0.03)	(-0.08, 0.05)	(-0.16, 0.09)	(-0.12, 0.01)
Account Age	(-0.60, 0.47)	(-0.04, 0.28)	(-0.26, 0.05)	(-0.28, 0.06)	(-0.25, 0.23)	(-0.15, 0.15)
Moderator Status Intercept[0]	(-3.16, -0.04)	(-2.66, 0.08)	(-2.55, 0.19)	(-2.64, 0.10)	(-2.89, -0.07)	(-2.38, 0.37)
Moderator Status Intercept [1]	(-2.36, 0.763)	(-2.47, 0.32)	(-2.21, 0.51)	(-2.50, 0.49)	(-2.30, 0.61)	(-2.58, 0.20)

Table 5: Regression coefficients for the MRP model in the practice support task. Each cell contains a 95% CI for the estimated parameter.

Variable Name/Model	Mod Survey Comments	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
# Comments (General)	(-0.28, 1.20)	(-0.14, 0.14)	(-0.20, 0.08)	(-0.09, 0.19)	(-0.16, 0.31)	(-0.19, 0.12)
# Comments (CMV)	(-0.18, 1.07)	(-0.03, 0.32)	(-0.00, 0.37)	(-0.11, 0.22)	(-0.22, 0.35)	(0.03, 0.50)
# Removals (General)	(-1.31, 0.16)	(-0.32, -0.04)	(-0.40, -0.06)	(-0.24, 0.09)	(-0.26, 0.26)	(-0.48, -0.10)
# Removals (CMV)	(-0.62, 0.63)	(-0.10, 0.03)	(-0.08, 0.06)	(-0.06, 0.09)	(-0.21, 0.07)	(-0.17, -0.01)
Account Age	(-0.40, 0.82)	(-0.23, 0.30)	(-0.10, 0.24)	(-0.30, 0.08)	(-0.36, 0.17)	(-0.07, 0.30)
Moderator Status Intercept[0]	(-2.93, 0.43)	(-2.74, 0.22)	(-2.61, 0.26)	(-3.14, -0.17)	(-3.153, -0.09)	(-2.85, -0.08)
Moderator Status Intercept [1]	(-3.17, 0.43)	(-2.60, 0.36)	(-2.53, 0.39)	(-2.38, 0.62)	(-2.42, 0.75)	(-2.20, 0.68)

Table 6: Regression coefficients for the MRP model in the rule application task. Each cell contains a 95% CI for the estimated parameter.

Variable Name/Model	Mod Survey Comments	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
# Comments (General)	(-0.35, 0.97)	(-0.09, 0.16)	(-0.19, 0.18)	(-0.11, 0.11)	(-0.17, 0.12)	(-0.20, 0.08)
# Comments (CMV)	(-0.97, 0.06)	(-0.01, 0.32)	(-0.21, 0.16)	(-0.17, 0.12)	(-0.10, 0.26)	(-0.04, 0.40)
# Removals (General)	(-0.92, 0.31)	(-0.27, -0.00)	(-0.30, 0.01)	(-0.13, 0.11)	(-0.33, -0.01)	(-0.27, 0.02)
# Removals (CMV)	(-0.597, 0.149)	(-0.278, 0.074)	(-0.347, 0.027)	(-0.396, 0.012)	(-0.241, 0.192)	(-0.257, 0.128)
Account Age	(-0.90, 0.16)	(-0.14, 0.16)	(-0.11, 0.20)	(-0.14, 0.18)	(-0.10, 0.26)	(-0.21, 0.14)
Moderator Status Intercept[0]	(-2.17, 1.01)	(-1.54, 1.16)	(-1.71, 1.41)	(-1.75, 1.15)	(-1.82, 1.28)	(-1.56, 1.59)
Moderator Status Intercept [1]	(-1,77, 1.32)	(-1.87, 0.84)	(-0.539, 1.733)	(41)	(-2.42, 0.77)	(-1.76, 1.51)

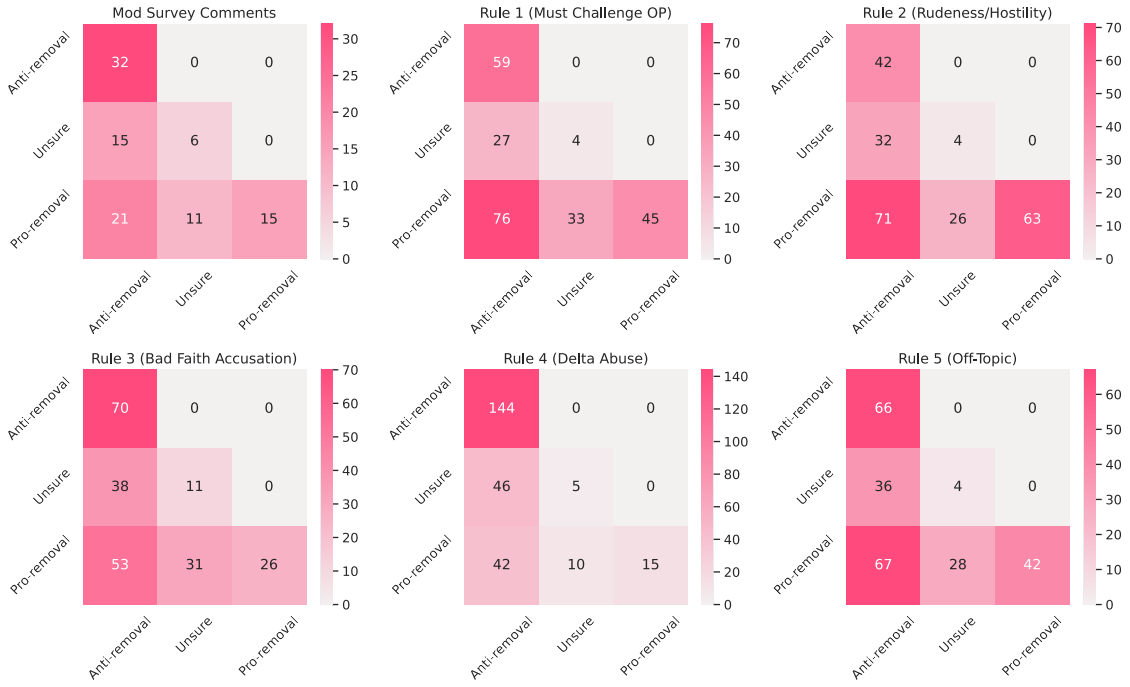

Figure 2: Co-occurrences of practice-awareness ratings by rule. The subplot in the top left corresponds to the set of comments that were also included in the survey sent to moderators.

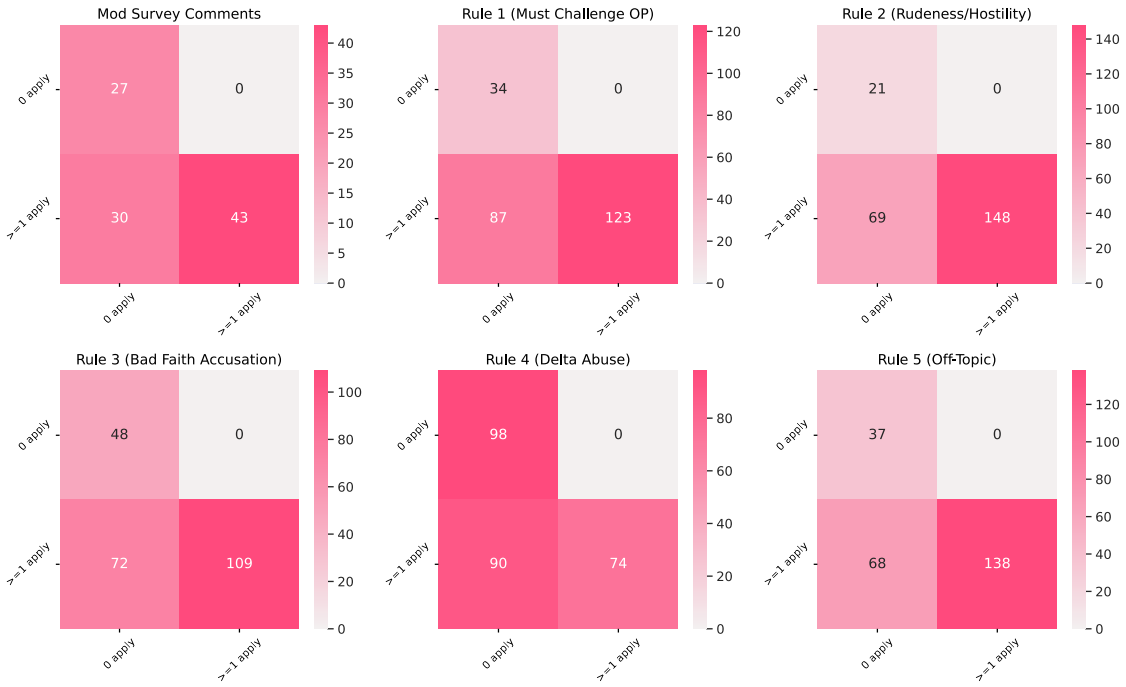

Figure 3: Co-occurrences of rule application ratings by rule.