

Does Transparency in Moderation Really Matter?

User Behavior after Content Removal Explanations on Reddit

Shagun Jhaver, Amy Bruckman, Eric Gilbert

Content Moderation

- Processes that determine which posts are allowed
- Black-box
- Difficult to form accurate mental model of content curation
- Lack of transparency => Decrease users' trust

Improving Transparency -> Explanations

Potential Role of Explanations

- Help users understand their mistakes
- Provide instructions on how to complete certain tasks
- Cost: Someone must spend time delivering explanations

Should moderators provide removal
explanations?

Main Contribution

- Evidence that offering explanations can help improve user behaviors
- Moderation should also be educational, not just punitive



Why Reddit?

- More than a million subcommunities
- Each subreddit has independent moderation
- Rich site for studying the diversity of explanations and effects

Research Questions

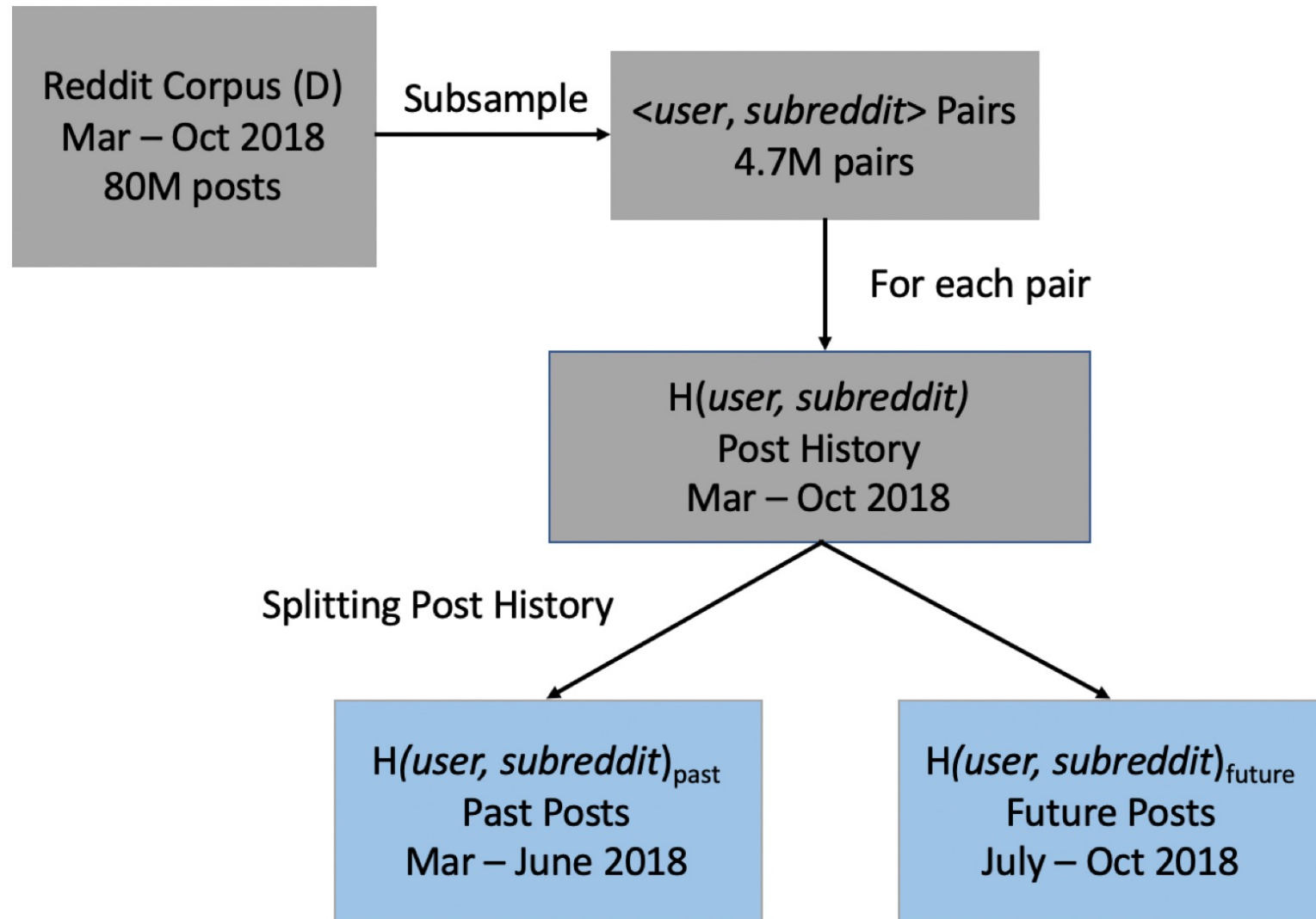
- What types of post removal explanations are typically provided to users?
- How does providing explanations affect the future posting activity of users?
- How does providing explanations affect the future post removals?

Related Work

- Impact of content moderation on end-users
 - Seering et al., Lampe et al.
- Evaluating legitimacy based on impact of human rights values
 - Suzor et al., Sloval et al., West
- Use of automated tools for content moderation
 - Long et al., Geiger & Ribers, Asthana & Halfaker

Data Preparation

Focus: Reddit submissions



Data

- 32 Million Submissions
- Extracted removal explanations for removed posts:
 - 213K explanation messages

Explanation Comment Example



[–] **AutoModerator** [M] 1 point 1 year ago

Hello, Your post has been removed because it breaks Rule 4a of [/r/mildlyinteresting](#), no videos. Note that only original photographs that you have taken yourself are allowed on this subreddit. As a result, this counts as a strike against your account. Three strikes will result in a ban. Please read the sidebar (hover over each rule) and [contact the mods](#) with a link to this post if you feel this was wrongfully removed.

I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](#) if you have any questions or concerns.

[permalink](#) [embed](#) [save](#) [give award](#)

RQ 1: What types of post removal explanations
are typically provided to users?

Topic Modeling of Explanation Comments

- LDA (Latent Dirichlet Allocation)
- Each explanation as a document
- Chose k , no. of topics, based on perplexity
 - $k = 28$

LDA Results (topic frequency in brackets)

- Reason why removal occurred:
 - e.g., Low karma (6.48%)
- Cushioning against dissatisfaction from removal:
 - e.g., Removal is unfortunate (4.93%)
- Educating beyond specific post:
 - e.g., Check rules in the sidebar (4.24%)

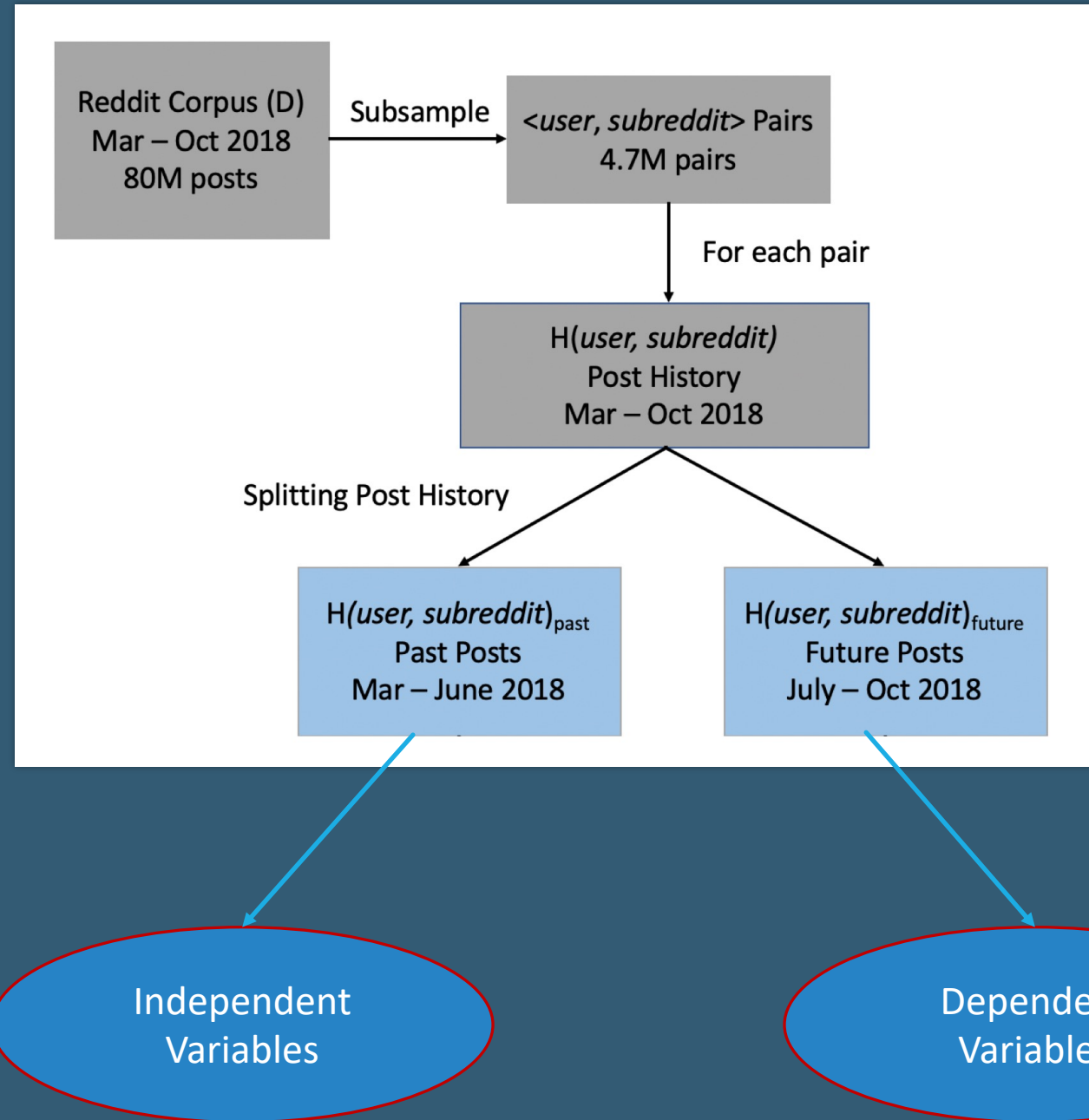
Who Provides Explanations?

- Human moderators.
- Automated Tools.
 - 58% of all explanations.

RQ 2: How does providing explanations affect the future posting activity of users?

RQ 3: How does providing explanations affect the future post removals?

Binomial Regression Models



Dependent Variables

○ Future Submission (RQ 2)

- Binary variable.

- For each $\langle u, s \rangle$ pair, whether the user u has a submission in s in the future.

○ Future Removal (RQ 3)

- Binary variable.

- For each $\langle u, s \rangle$ pair, whether the user u has a submission in s in the future that was removed.

Control Variables: Subreddit Variables

- # Subreddit Subscribers
- # Subreddit Submissions
- Net Subreddit Removal Rate

Control Variables: Post History Variables

- Past Submissions by user u in subreddit s
- Avg. Past Score
- Avg. Past Comments

Independent Variables

- Past Removal Rate for (u, s)
- Explanation Rate
- Explanation by Bot Rate

Output Variable	Model	Input variables	Inclusion criteria	Valid entries
Future Submission	A.1	Subreddit variables + Post history variables + Past Removal Rate	All <user, subreddit> pairs	4.7M
	A.2	+ Explanation Rate	Past Removal Rate > 0	1.4M
	A.3	+ Average Explanation Length + Explanation through Comments Rate	Explanation Rate > 0	147.8K
	A.4	+ Explanation by Bot Rate	Explanation through Comments Rate > 0	31K
Future Removal	B.1	Subreddit variables + Post history variables + Past Removal Rate	Future Submissions > 0	1.8M
	B.2	+ Explanation Rate	Future Submissions > 0 AND Past Removal Rate > 0	548.7K
	B.3	+ Average Explanation Length + Explanation through Comments Rate	Future Submissions > 0 AND Explanation Rate > 0	64.8K
	B.4	+ Explanation by Bot Rate	Future Submissions > 0 AND Explanation through Comments Rate > 0	15.2K

Findings: Future Submission

Reporting the results

- Odds ratios percentage

- Percentage change in the odds of DV when an IV is increased by one SD

Observations

Independent Variable (1 SD increase)	Posting in the future
Past Removal Rate	2.2% lower odds***

Higher past removals ~ Lower future posts

Observations

Independent Variable (1 SD increase)	Posting in the future
Past Removal Rate	2.2% lower odds***
Explanation Rate	1.2% lower odds***

Higher explanations ~ Lower future posts

Observations

Independent Variable (1 SD increase)	Posting in the future
Past Removal Rate	2.2% lower odds***
Explanation Rate	1.2% lower odds***
Explanation by Bot Rate	26.4% higher odds***

Higher explanations by bots ~ Higher future posts

Findings: Future Removal

Observations

Independent Variable (1 SD increase)	Post removals in the future
Past Removal Rate	96.8% higher odds***

Higher past removals ~ Higher future removals

Observations

Independent Variable (1 SD increase)	Post removals in the future
Past Removal Rate	96.8% higher odds***
Explanation Rate	6.5% lower odds***

Higher explanations ~ Lower future removals

Observations

Independent Variable (1 SD increase)	Post removals in the future
Past Removal Rate	96.8% higher odds***
Explanation Rate	6.5% lower odds***

Good News for Moderators who provide explanations!

Observations

Independent Variable (1 SD increase)	Post removals in the future
Past Removal Rate	96.8% higher odds***
Explanation Rate	6.5% lower odds***
Explanation by Bot Rate	No sign. effect

Nagelkerke R Square = 0.231

Takeaways

Removal Explanations Help Users Learn Social Norms

- Direct feedback
- Which rule is broken
- Public posting
- Linked to reduction in post removals
 - Reduced workload for moderators
- Moderation should also play an educational role

How should Removal Explanations be provided?: Human moderators versus automated tools

- Human mods have no advantage for reducing future post removals
- Opportunity for deploying tools for explanations
 - Explainable AI

Does Transparency in Moderation Really Matter?

User Behavior after Content Removal Explanations on Reddit

Shagun Jhaver, Amy Bruckman, Eric Gilbert



@shagunjhaver



jhaver.shagun@gatech.edu



Regression Results – Future Submissions

Group	Variables	Model A.1	Model A.2	Model A.3	Model A.4
Subreddit variables	Subreddit Subscribers	0.968***	0.988***	0.984*	0.981
	Subreddit Submissions	1.118***	1.164***	1.146***	1.151***
	Net Subreddit Removal Rate	0.940***	0.938***	0.900***	0.86***
Post history variables	Past Submissions	7.4E+10***	4.5E+6***	687.6***	16.628***
	Average Past Score	0.995***	0.999	0.988	0.972
	Average Past Comments	1.037***	1.014**	1.043***	1.057**
Independent variables	Past Removal Rate	0.978***	0.638***	0.563***	0.520***
	Explanation Rate		0.988***	0.686***	0.636***
	Average Explanation Length			1.003	0.990
	Explanation through Comments Rate			1.035***	0.824***
	Explanation by Bot Rate				1.264***
	# Obs	4.7M	1.4M	147.8K	31K
	Intercept	1.135***	1.154***	1.218***	1.355***
	Nagelkerke R Square	0.191	0.267	0.337	0.327
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

Regression Results – Future Removals

Group	Variables	Model B.1	Model B.2	Model B.3	Model B.4
Subreddit variables	Subreddit Subscribers	0.910***	0.934***	0.891***	0.93**
	Subreddit Submissions	1.168***	1.033***	1.11***	1.079**
	Net Subreddit Removal Rate	2.461***	2.215***	2.058***	2.021***
Post history variables	Past Submissions	1.164***	2.6***	5.636***	2.443***
	Average Past Score	0.991***	0.981***	0.96**	0.975
	Average Past Comments	1.02***	1.000	1.027	1.0
Independent variables	Past Removal Rate	1.968***	1.366***	1.236***	1.286***
	Explanation Rate		0.935***	0.701***	0.649***
	Average Explanation Length			1.003	1.002
	Explanation through Comments Rate			0.905***	0.774***
	Explanation by Bot Rate				1.019
	# Obs	1.8M	548.7K	64.8K	15.2K
	Intercept	0.392***	2.148***	2.287***	2.044***
	Nagelkerke R Square	0.378	0.187	0.199	0.231
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

Subreddit Models


Subreddit	r/politics		r/pics		r/mildlyinteresting		r/buildapc	
Dependent Var.	Future subm.	Future removal	Future subm.	Future removal	Future subm.	Future removal	Future subm.	Future removal
Past Submissions	5022 ***	13.682 ***	2.181 ***	1.217 ***	3.108 ***	1.778 **	1.119	1.048
Avg Past score	0.972	0.889	1.287	1.229	1.12	0.902	0.684	1.133
Avg Past comments	1.04	1.139	0.846	0.834	0.926	1.28	1.56	1.224
Past Removal Rate	0.538 ***	1.521 ***	0.586 ***	1.756 ***	0.771 ***	1.36 ***	0.779 ***	1.331 **
Explanation Rate	0.997	0.877 *	0.943	0.561 ***	1.02	0.795 ***	0.774 *	0.48 ***
Intercept	7.304 ***	7.367 ***	1.317	0.481 ***	0.536 ***	0.431 ***	2.329 *	0.052 ***
Nag. R Square	0.382	0.084	0.204	0.21	0.104	0.049	0.188	0.224
# Obs	4357	2537	4105	1404	4670	1368	419	174

Explanation Flair Example

↑



4666

↓



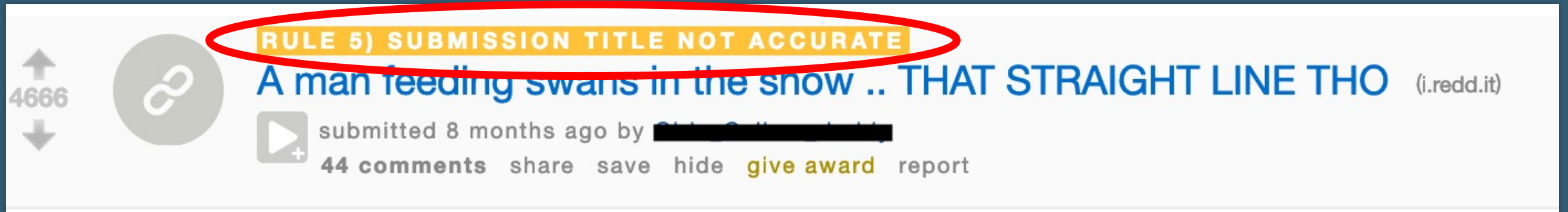
RULE 5) SUBMISSION TITLE NOT ACCURATE

A man feeding swans in the snow .. THAT STRAIGHT LINE THO (i.redd.it)

 submitted 8 months ago by 

44 comments share save hide **give award** report

Explanation Flair Example



Explanation Flairs

Unigram		Bigram		Trigram	
Phrases	Frequency	Phrases	Frequency	Phrases	Frequency
removed	65817	removed rule	20149	non whitelisted domain	2671
rule	53902	low karma	7038	rule overposted content	2252
fluff	24773	removed repost	3119	rule non gore	1350
repost	17485	fluff question	2896	r14 social media	1179
low	10737	non whitelisted	2671	social media sms	1179
submitted	10238	whitelisted domain	2671	media sms removed	1179
karma	7138	low effort	2593	removed crappy design	1151
title	6816	repost removed	2548	use approved host	1110
content	5406	rule overposted	2252	approved host removed	1110
post	4397	overposted content	2252	assign flair post	979
non	4299	appropriate subreddit	1629	low effort meme	902
shitpost	3812	rule repost	1602	removed restricted content	875
question	3774	social media	1594	removed location missing	849
domain	3387	rule animeme	1528	removed low quality	715
r1	3254	rule non	1491	r3 repost removed	637

For each (user u , subreddit s):

Moderation decisions + removal explanations for
prior posts by u in s



Future posting behavior
by u in s