Can we better manage conversations on Twitter, Facebook, Reddit, etc.?
Something's brewing! Early prediction of controversy-causing posts from discussion features

Looks like the proposal will be controversial.

Jack Hessel and Lillian Lee, NAACL 2019

#DSRD19
Task: predict whether a social media post will get many positive and negative responses, or no?

Yes, controversial

No, not controversial

#DSRD19
Utility to site moderators and administrators

Controversy (as we have defined it) is not necessarily a bad thing.

- Monitoring for “bad” controversy can prevent harm to the group
- Bringing “productive” controversy to the community’s attention can help the group solve problems
Observation: controversy is community-specific

“break up”: controversial in the Reddit group on relationships, but not in the group for posing questions to women

“my parents”: controversial for the personal-finance group (example: “live with my parents”) but not in the relationships group
Our datasets ("fill-in" of Baumgartner's crawl)

- 6 communities on [www.reddit.com](http://www.reddit.com):
  - two QA subreddits: *AskMen*, *AskWomen*
  - a special interest community: *Fitness*
  - three advice communities: *LifeProTips*, *personalfinance*, *relationships*
- Posts and comments mostly web-English
- Up/downvote information: *eventual* percent-upvoted
  (we can’t use early votes: no timestamps)

#DSRD19
Observation: we can use early reactions

• Early opinions can greatly affect subsequent opinion dynamics (Salganik et al. MusicLab experiment, Science 2006, inter alia)

• Both the content and structure of the early discussion tree may prove helpful.
Early comments: how many?

=15% of eventual =32% of eventual

#DSRD19
We predict community-specific controversy of a post, examining domain transferability of features, using an early detection paradigm.

Predicting controversy from posting-time-only features
(Dori-Hacohen and Allan, 2013; Mejova et al., 2014; Klenner et al., 2014; Dori-Hacohen et al., 2016; Jang and Allan, 2016; Jang et al., 2017; Addawood et al., 2017; Timmermans et al., 2017; Rethmeier et al., 2018; Kaplun et al., 2018)

Retrospective analyses: was a given hashtag/entity/word controversial previously?
(Popescu and Pennacchiotti, 2010; Choi et al., 2010; Rad and Barbosa, 2012; Cao et al., 2015; Lourentzou et al., 2015; Chen et al., 2016; Addawood et al., 2017; Beelen et al., 2017; Al-Ayyoub et al., 2017; Garimella et al., 2018)

Disagreement or antisocial behavior
(Mishne and Glance, 2006; Yin et al., 2012; Awadallah et al., 2012; Allen et al., 2014; Wang and Cardie, 2014; Marres, 2015; Borra et al., 2015; Jang et al., 2017; Basile et al., 2017; Liu et al., 2018; Zhang et al., 2018; Chang & Danescu-Niculescu-Mizil., 2019)
Prediction results incorporating comment features: One community

AskWomen

4 comments, on average

Best baseline on original post: Meanpool BERT 1st 512 words, L2 normalize, PCA-> 100 dims, linear classifier

*Significant diff over baseline at 45 mins

#DSRD19
Tree/Rate features transfer better than content

Testing Subreddit

Training Subreddit

(a) TEXT+RATE+TREE
\[ t = 180 \]

(b) RATE+TREE
\[ t = 180 \]

#DSRD19
Takeaways (modulo caveats! see paper)

• We advocate an early-detection, community-specific approach to controversial-post detection
  
  ○ Early detection outperforms posting-time-only features in 5 of 6 Reddit communities tested, even, sometimes, for quite small early-time windows

  ○ Early comment content is most effective, but tree-shape and rate features transfer across domains better
Content removal as a moderation strategy

What if I censure that "oink"?

rule-breaking comment

Kumar Bhargav Srinivasan
Cristian Danescu-Niculescu-Mizil
Lillian Lee
Chenhao Tan

CSCW 2019

#DSRD19
Test case: **ChangeMyView subreddit:** Known to be surprisingly productive

- CMV moderators manually removed **22,788** comments between January 2015 and March 2018.

- Users consider moderator intervention to be one of the main factors behind the quality of discussions in CMV.
  - “I’ve seen threads go ugly so fast [on other subreddits], and I think that having active mods helps CMV not get bogged down by trolls.”
  
  [Jhaver, Vora, Bruckman 2017]

- We have moderator-log access through previous CMV work.

#DSRD19
comment deletion for rule violation on CMV

Comment removed by moderator 4 months ago

[username]
your comment has been removed for breaking Rule 2:

Don't be rude or hostile to other users. Your comment will be removed even if most of it is solid, another user was rude to you first, or you feel your remark was justified. Report other violations; do not retaliate. See the wiki page for more information.

If you would like to appeal, message the moderators by clicking this link. Please note that multiple violations will lead to a ban, as explained in our moderation standards.

#DSRD19
Comment deletion and future activity (or lack thereof)
The effect of comment deletion on those who stay?

Possible reasons that comment deletion may not cause compliant behavior:


- (and see two slides from now)
In this work, we don't do A/B testing

• Randomizing comment deletion may disrupt a popular and productive community.

• Randomizing comment removals seems wrong for non-violating comments.
Interrupted time-series analysis at removal?

"Comment 0" made, then deleted by mod

Stat. sig change in slope, level?

Confound: effect of having made a removal-meriting comment.
(Drop an "F-bomb", then self-censor regardless of moderator action?)

8.4K discussion trees with total 22K mod-removed comments, 73K trees and 4M comments total
Observational delayed-feedback paradigm

Delay (>2 hours in 40% of cases)

User's comment timeline

Comment that will be removed is made

That comment is removed by a mod

#DSRD19
Delayed-feedback paradigm

Delay (pre-removal window)

If $c_{-1}$ is not rule-abiding, but $c_1$ is, now do we know deletion is the cause?

Alas, no – cannot rule out temporal effects.

#DSRD19
Delayed-feedback paradigm

"treatment" user

"control" user

Delay (pre-removal window)

Matched delay

#DSRD19
Less non-compliance (non-targe-deletion trees)?

\[ \beta_2 = -1.31 \times 10^0 (***) \]
\[ \beta_3 = -8.35 \times 10^{-2} (***) \]

- Intermittent time series
- Delayed feedback

\#DSRD19
Increased engagement (comment length)?

\[
\begin{align*}
\beta_2 &= 6.22e+0^{(*)} \\
\beta_3 &= 9.18e-2
\end{align*}
\]

Interrupted time series

Delayed feedback

DF treatment

DF control

before \((c_{-1} \text{ or } c'_{-1})\)

after \((c_1 \text{ or } c'_1)\)

#DSRD19
Takeaways (modulo caveats! see paper)

- "Delayed feedback" observational paradigm – better controls compared to "standard" ITS application
  - Limitation: only applicable to users active enough to post in the delay window
- For applicable users, comment moderator-deletion causes immediate non-compliance drop with no significant change in "post effort" (length)
Summary: "Movie trailers" of controversy, comment removal

Please see the NAACL 2019 and CSCW 2019 paper for (many more) details: http://www.cs.cornell.edu/home/llee/papers.html

Thanks for listening!