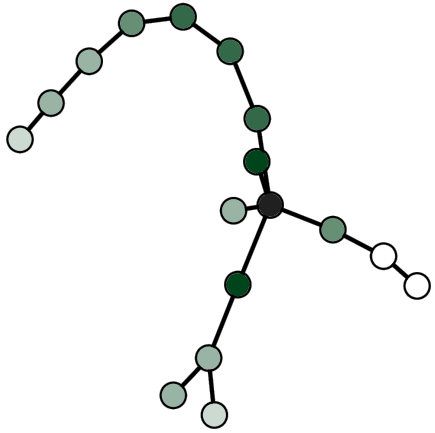
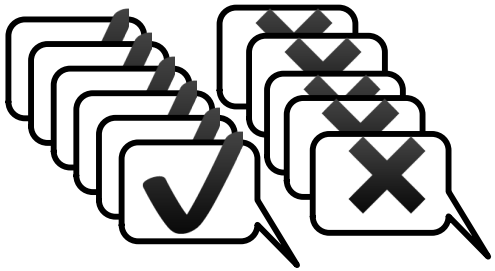


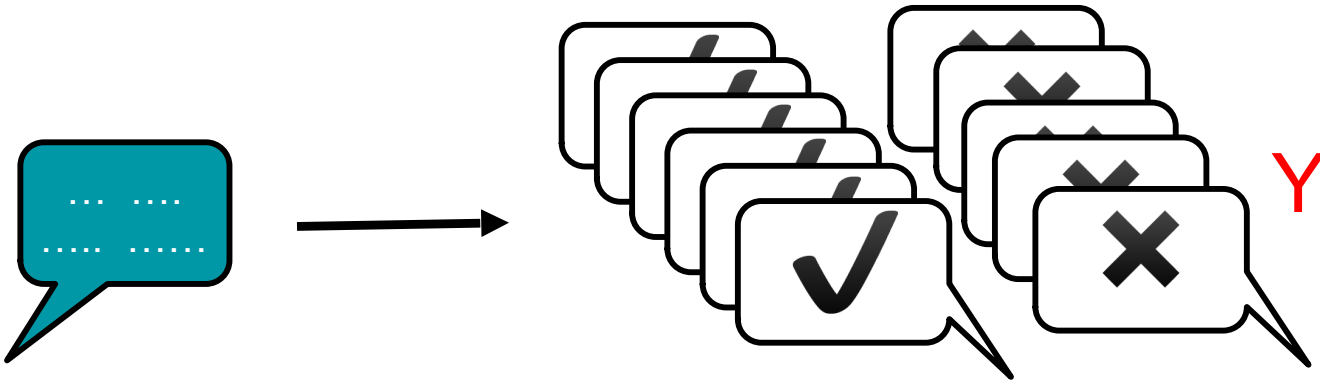
# Something's brewing!

## Early prediction of controversy-causing posts from discussion features



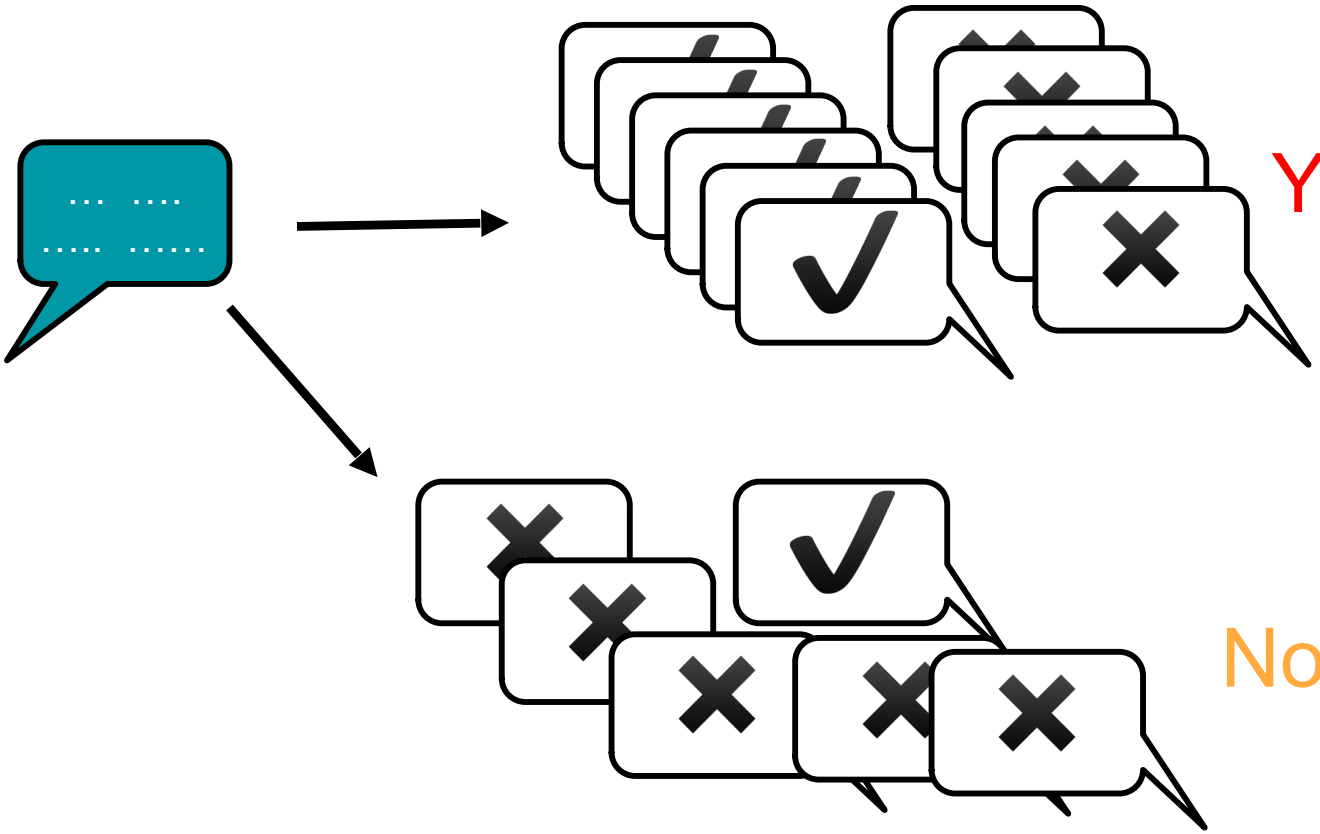
Jack Hessel and Lillian Lee
Cornell University

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



Yes, **controversial**

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



Yes, **controversial**

No, not controversial

## 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 personal-finance group
(example: "live with my parents")
but not in the relationships group

# Observation: we can also 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.

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

**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)

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

**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; Zhang et al., 2018)

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

**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)

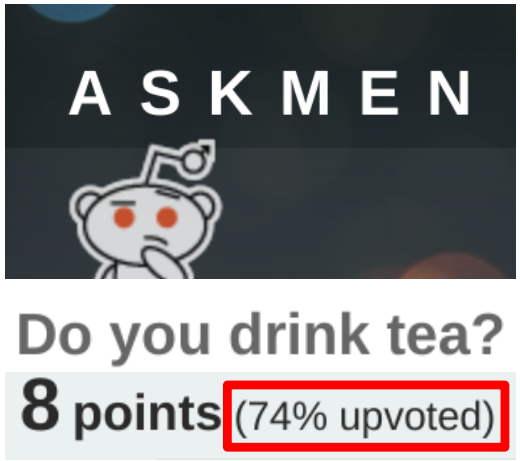
**Disagreement or antisocial behavior**
(Mishne and Glance, 2006; Yin et al., 2012; Awadallah et al., 2012; Allen et al., 2014; Wang and Cardie, 2014; Marres, 2015; Borra et al., 2015; Jang et al., 2017; Basile et al., 2017; Liu et al., 2018; Zhang et al., 2018; Zhang et al., 2018)

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

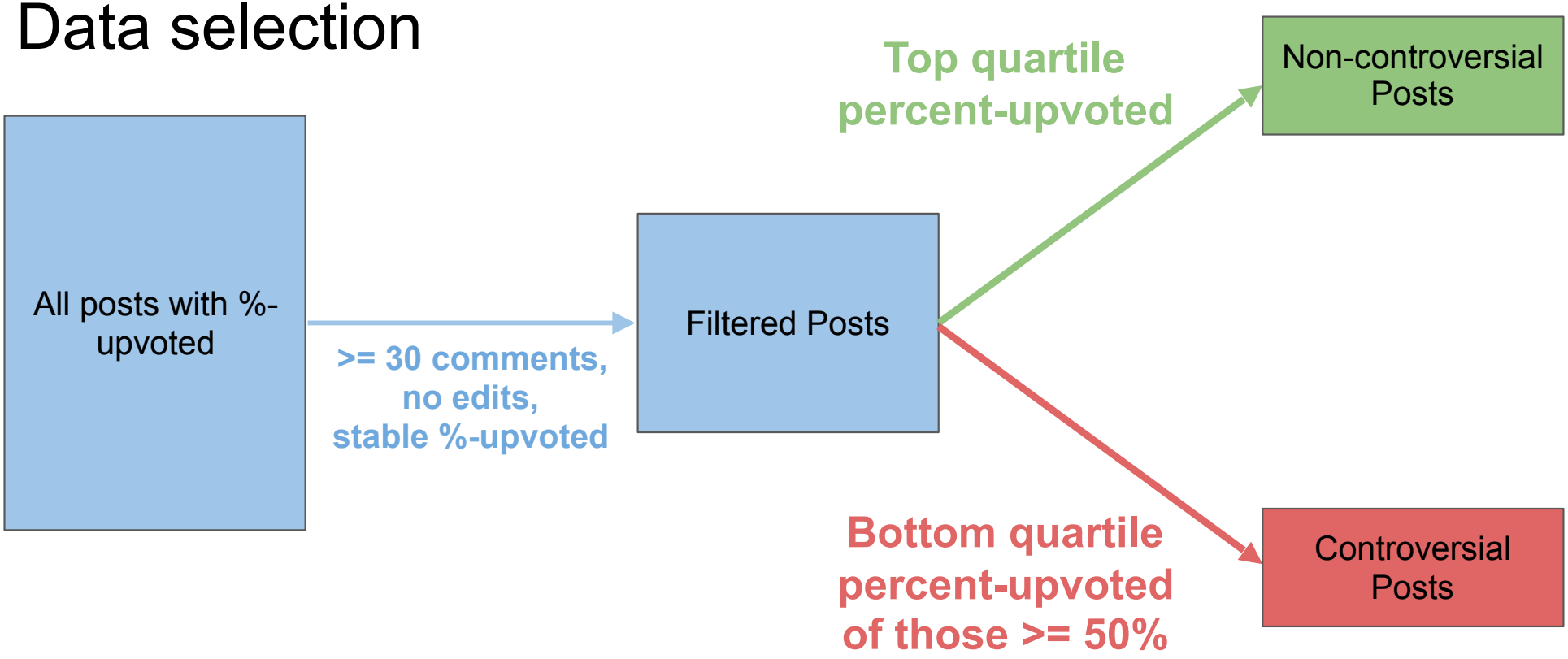
**Predicting controversy from posting-time-only features**
(Dori-Hacohen and Allan, 2013; Mejova et al., 2014; Klenner et al., 2014; Dori-Hacohen et al., 2016; Jang and Allan, 2016; Jang et al., 2017; Addawood et al., 2017; Timmermans et al., 2017; Rethmeier et al., 2018; Kaplun et al., 2018)

# Our datasets (derived from Baumgartner)

- 6 communities on 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)



# Data selection




Label validation steps (details in paper):

1) high-precision overlap (>88 F-measure) with reddit's low-recall rank-by-controversy

2) we ensure popularity prediction != controversy prediction

# Labeled Dataset Statistics

| | # posts | # cmnts | $\mu_{up}$ cont | $\mu_{up}$ noncont |
|---|---|---|---|---|
| AskMen | 3.3K | 474K | 66% | 90% |
| AskWomen | 3.0K | 417K | 67% | 91% |
| Fitness | 3.9K | 625K | 66% | 91% |
| LifeProTips | 1.6K | 208K | 68% | 91% |
| personalfinance | 1.0K | 95K | 72% | 92% |
| relationships | 2.2K | 221K | 68% | 93% |

Balanced, binary classification with **controversial**/**non-controversial** labeling

Performance metric: accuracy

## Some posting-time-text-only results (this, plus timestamp, is our baseline)

| | AM | AW | FT | LT | PF | RL |
|---|---|---|---|---|---|---|
| HAND | 55.4 | 52.2 | 61.9 | 59.7 | 54.5 | 60.8 |
| TFIDF | 57.4 | 60.1 | 63.3 | 59.1 | 58.7 | 65.4 |
| ARORA | 58.6 | 62.0 | 60.5 | 59.4 | 57.2 | 62.1 |
| W2V | 60.7 | 62.1 | 63.1 | 61.4 | 59.9 | 64.3 |
| LSTM | 58.9 | 58.2 | 63.6 | 61.5 | 60.0 | 63.1 |
| BERT-LSTM | **64.5** | **65.1** | **66.2** | 65.0 | 65.1 | 67.8 |
| BERT-MP | 63.4 | 64.0 | 64.4 | 65.7 | 64.1 | 67.0 |
| BERT-MP-512 | 63.9 | 64.0 | 64.7 | **65.8** | **65.6** | 67.7 |
| HAND+W2V | 61.3 | 62.3 | 64.9 | 63.2 | 60.0 | 66.3 |
| HAND+BERTMP512 | 63.6 | 63.5 | 64.9 | 64.1 | 64.4 | **68.0** |

Table 2: Average accuracy for each post-time, text-only predictor for each dataset, averaged over 15 cross-validation splits; standard errors are $\pm.6$, on average (and never exceed $\pm 1.03$). Bold is best in column; underlined are statistically indistinguishable from best in column ($p < .01$)

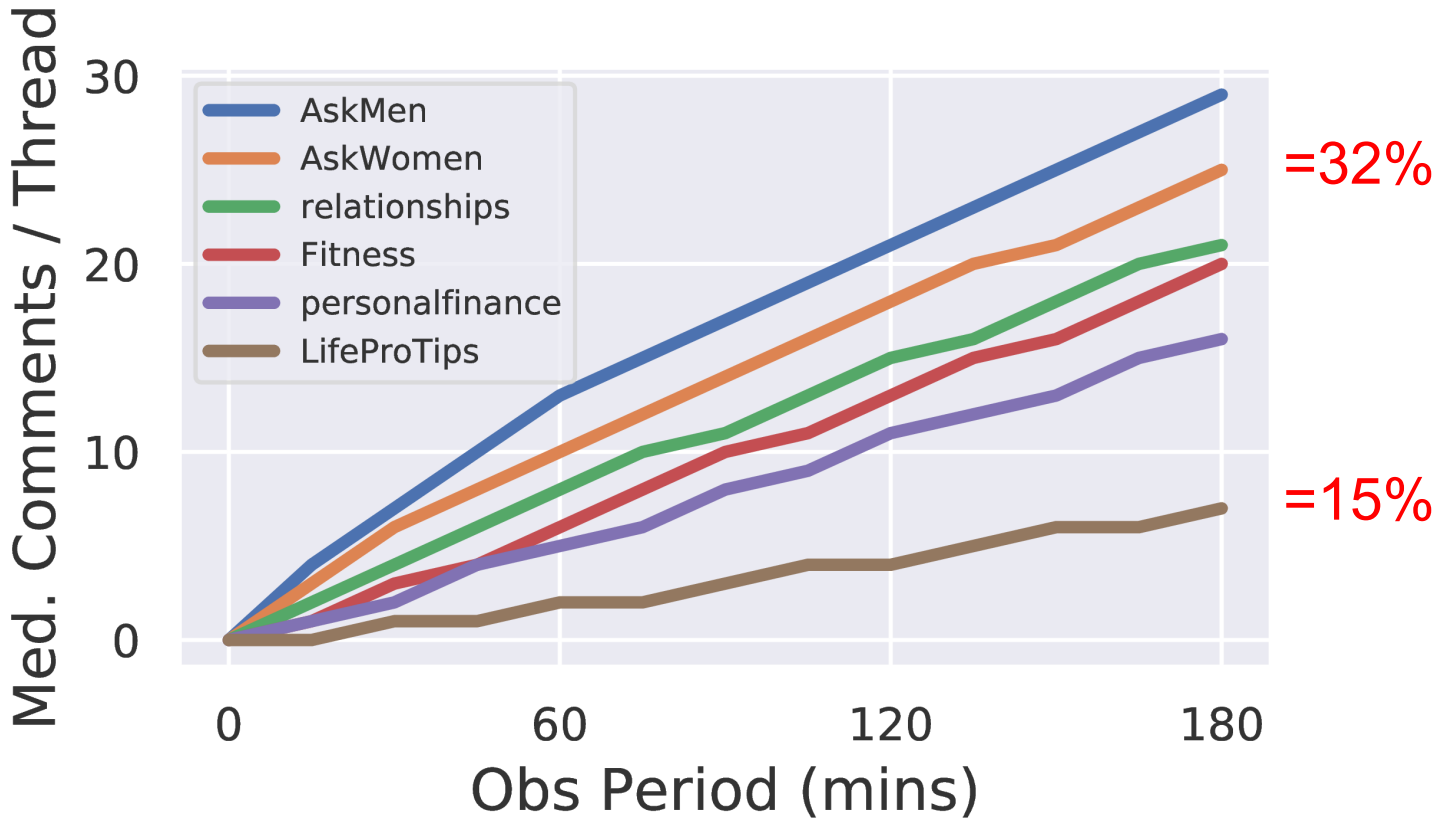
## Some posting-time-text-only results (this, plus timestamp, is our baseline)

| | AskMen | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| HAND-crafted | | | | | | |
| Word2Vec | | | | | | |
| W2V-LSTM | | | | | | |
| BERT-LSTM | ☆ | ☆ | ☆ | ○ | ○ | ○ |
| BERT-meanpool-512-then-linear | ○ | ○ | ○ | ☆ | ☆ | ○ |
| HAND+W2V | | | ○ | ○ | | ○ |
| HAND+BERT-meanpool-512 then linear | ○ | ○ | ○ | ○ | ○ | ☆ |

- Rather than passing BERT vectors to a bi-LSTM, it works about as well and faster to mean-pool, dimension-reduce, and feed to a linear classifier

- Our hand-crafted features + word2vec match BERT-based algorithms on 3 of 6 subreddits

# Early comments: how many?



# Does the shape of the tree predict controversy?

Usually yes, even after controlling for the rate of incoming comments.

## Tree Features

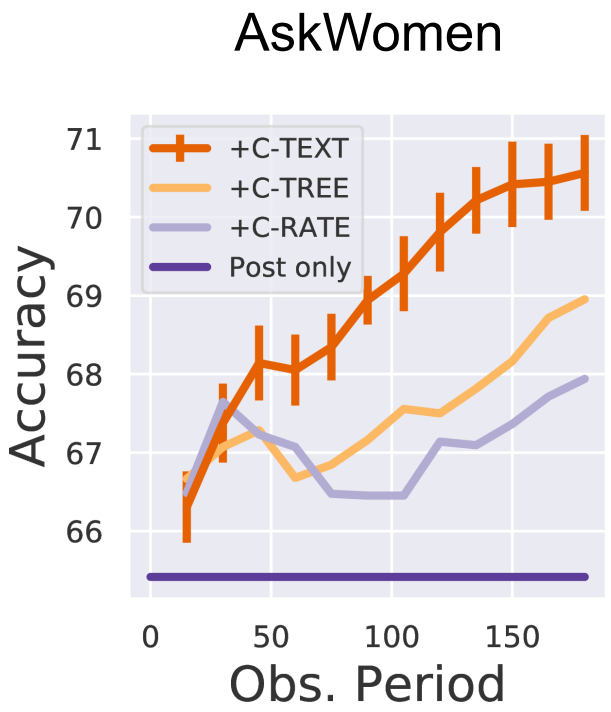
- max depth/total comment ratio
- proportion of comments that were top-level (i.e., made in direct reply to the original post)
- average node depth
- average branching factor
- proportion of top-level comments replied to
- Gini coefficient of replies to top-level comments (to measure how "clustered" the total discussion is)
- Wiener Index of virality (average pairwise pathlength between all pairs of nodes)

## Rate Features

- total number of comments
- logged time between OP and the first reply
- average logged parent-child reply time (over all pairs of comments)

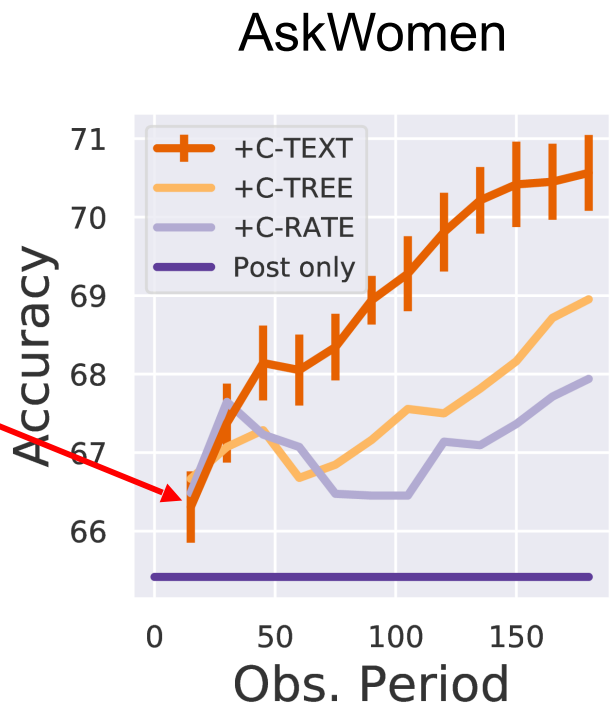
[binary logistic regression, LL-Ratio test $p<.05$ in 5/6 communities]

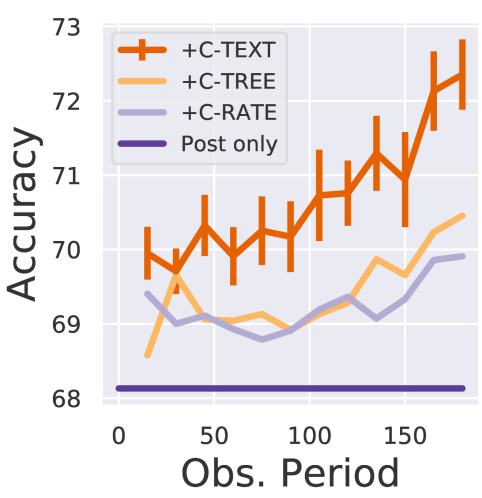
# Prediction results incorporating comment features



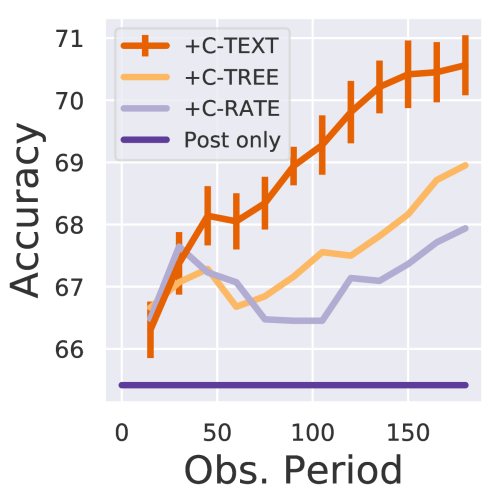
# Prediction results incorporating comment features

4 comments,
on average

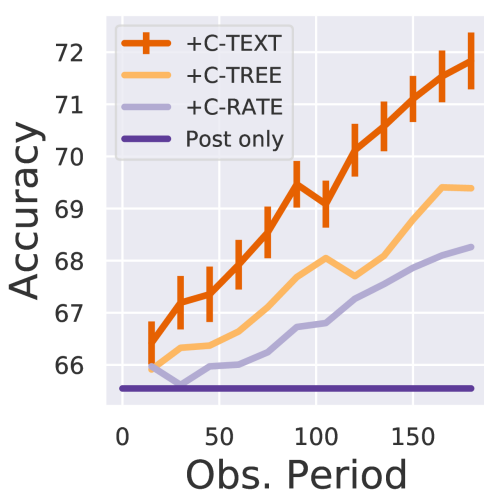





AskMen

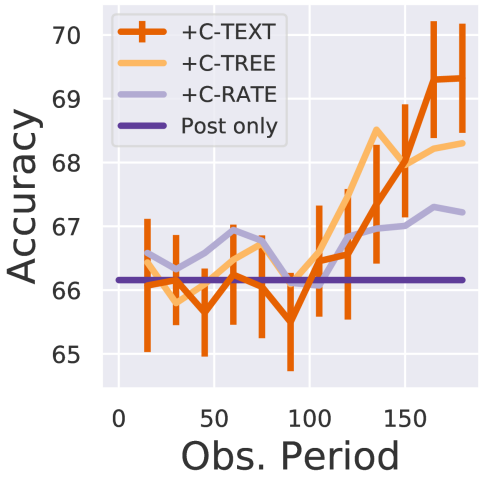



AskWomen

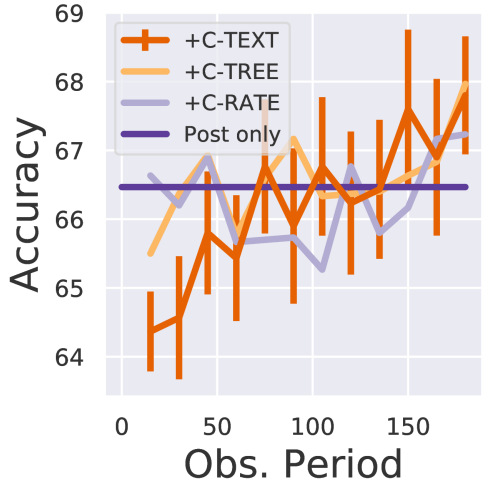



Fitness

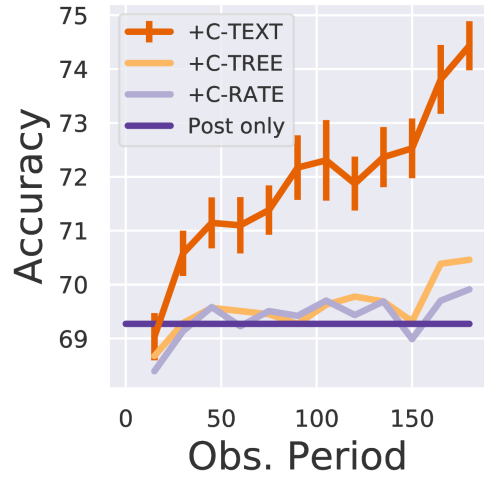
LifeProTips



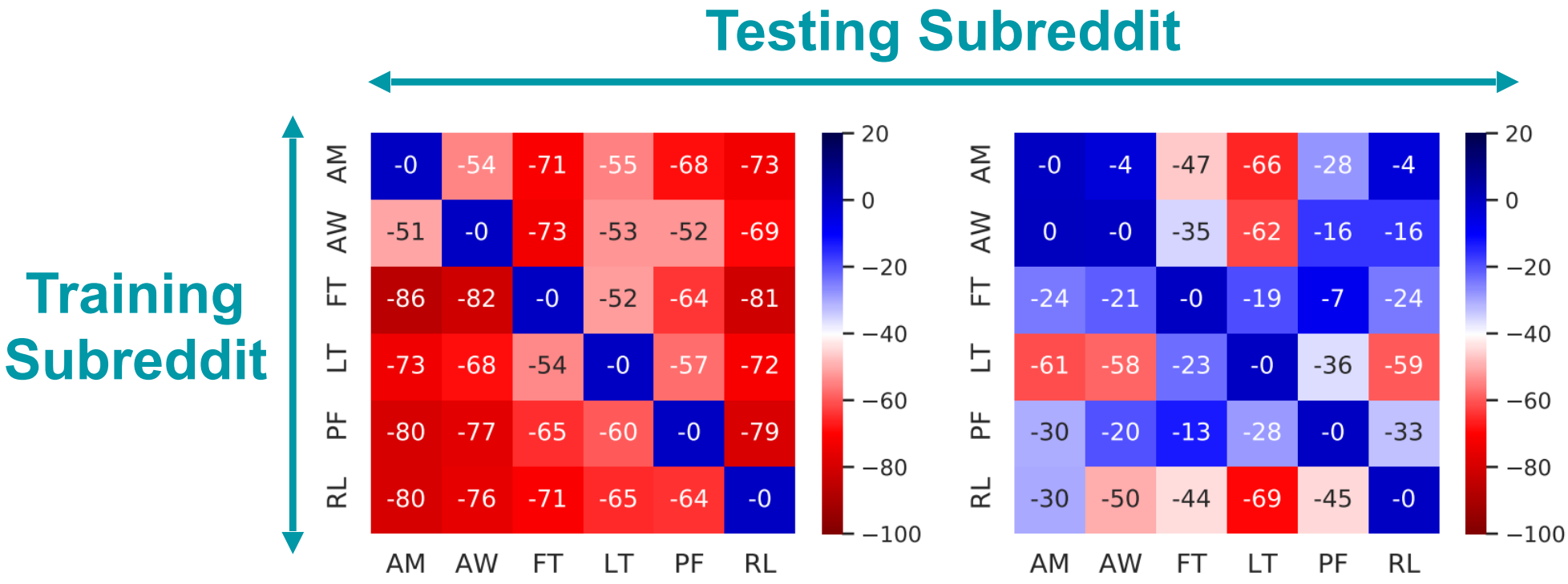

personalfinance




relationships



# Tree/Rate features transfer better than content




(a) TEXT+RATE+TREE $t = 180$

(b) RATE+TREE $t = 180$

## Takeaways (modulo caveats! see paper)

- We advocate an early-detection, community-specific approach to controversial-post prediction
  - We can use features of the content and structure of the early discussion tree
  - Early detection outperforms posting-time-only features in 5 of 6 Reddit communities tested, even for quite small early-time windows
  - Early content is most effective, but tree-shape and rate features transfer across domains better