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?
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.
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.
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)

**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)

**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)
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

<table>
<thead>
<tr>
<th></th>
<th># posts</th>
<th># cmnts</th>
<th>$\mu_{up}$ cont</th>
<th>$\mu_{up}$ noncont</th>
</tr>
</thead>
<tbody>
<tr>
<td>AskMen</td>
<td>3.3K</td>
<td>474K</td>
<td>66%</td>
<td>90%</td>
</tr>
<tr>
<td>AskWomen</td>
<td>3.0K</td>
<td>417K</td>
<td>67%</td>
<td>91%</td>
</tr>
<tr>
<td>Fitness</td>
<td>3.9K</td>
<td>625K</td>
<td>66%</td>
<td>91%</td>
</tr>
<tr>
<td>LifeProTips</td>
<td>1.6K</td>
<td>208K</td>
<td>68%</td>
<td>91%</td>
</tr>
<tr>
<td>relationships</td>
<td>1.0K</td>
<td>95K</td>
<td>72%</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td>2.2K</td>
<td>221K</td>
<td>68%</td>
<td>93%</td>
</tr>
</tbody>
</table>

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

Performance metric: accuracy
Some posting-time-text-only results (this, plus timestamp, is our baseline)

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

Table 2: Average accuracy for each post-time, text-only predictor for each dataset, averaged over 15 cross-validation splits; standard errors are ±.6, on average (and never exceed ±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)

<table>
<thead>
<tr>
<th></th>
<th>AskMen</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAND-crafted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word2Vec</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2V-LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-LSTM</td>
<td>⭐⭐⭐⭐⭐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-meanpool-512-then-linear</td>
<td>⬜ ⬜ ⬜ ⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAND+W2V</td>
<td>⬜ ⬜ ⬜</td>
<td>⬜</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAND+BERT-meanpool-512 then linear</td>
<td>⬜ ⬜ ⬜ ⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
</tbody>
</table>

- 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?

- AskMen: 32% (at 180 mins)
- AskWomen: 15% (at 180 mins)
- relationships
- Fitness
- personalfinance
- LifeProTips

Graph showing the trend of median comments per thread over the observation period (in minutes) for different categories.
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

![Graph showing prediction results for AskWomen](image)
Prediction results incorporating comment features

4 comments, on average
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