POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection

Yujian Liu¹,*, Xinliang Frederick Zhang¹,*, David Wegsman¹, Nick Beauchamp², and Lu Wang¹

¹Computer Science and Engineering, University of Michigan,
²Political Science and Network Science Institute, Northeastern University
*Equal contributions

NAACL 2022
Analyze political text

Conservatives continue to support baseless claims of election fraud ...
Trump isn’t the only Republican making unsubstantiated claims about voter fraud, a move that may put lives at risk.

I don’t *believe* life begins at conception--I *know* life begins at conception. #nrlc2015

Political ideology?

Conservatives

Liberal

Stance towards legal abortion?

I don’t *believe* life begins at conception--I *know* life begins at conception. #nrlc2015

Against

Semeval-2016 task 6: Detecting stance in tweets (Mohammad et al., 2016)
More than word choice…

- Selection of entities and events to present (Framing Theory; Entman, 1993)
- Key: consider the global context of the given article
- Approach: compare articles from different media outlets that report the same story

---

**News Story:** Donald Trump tests positive for COVID-19.

**Daily Kos** (left): It’s now clear that Donald Trump lied to the nation about when he received a positive test for COVID-19. ... they’re continuing to act as if nothing has changed—and that **disregarding science** and **lying** to the public are the only possible strategies.

**The Washington Times** (right): *Trump says he’s “doing very well” ... President Trump thanked the nation for supporting him* Friday night as he left the White House to be hospitalized for COVID-19. **“I want to thank everybody for the tremendous support...”** Mr. Trump said in a video recorded at the White House.

**Breitbart** (right): *President Donald Trump thanked Americans for their support* on Friday as he traveled to Walter Reed Military Hospital for further care after he was diagnosed with coronavirus. **“I think I’m doing very well...”** Trump said in a video filmed at the White House and posted to social media.
BIGNEWS dataset

- **BIGNEWS**: 3.7M US political news articles from 11 news outlets

<table>
<thead>
<tr>
<th>Ideology</th>
<th>Daily Kos</th>
<th>HPO</th>
<th>CNN</th>
<th>WaPo</th>
<th>NYT</th>
<th>USA Today</th>
<th>AP</th>
<th>The Hill</th>
<th>TWT</th>
<th>FOX</th>
<th>Breitbart</th>
</tr>
</thead>
<tbody>
<tr>
<td># articles</td>
<td>100,828</td>
<td>241,417</td>
<td>64,988</td>
<td>198,529</td>
<td>173,737</td>
<td>170,737</td>
<td>279,312</td>
<td>322,145</td>
<td>243,181</td>
<td>330,166</td>
<td>206,512</td>
</tr>
<tr>
<td># words</td>
<td>738.7</td>
<td>729.9</td>
<td>655.7</td>
<td>803.2</td>
<td>599.4</td>
<td>691.7</td>
<td>572.3</td>
<td>426.3</td>
<td>522.7</td>
<td>773.5</td>
<td>483.5</td>
</tr>
</tbody>
</table>

- **BIGNEWSALIGN**: 1M clusters of articles that report the same story
  - Aligned by text and entity similarities
Continued pretraining objectives

- Ideology objective: Acquire ideology-informed representations

\[ \mathcal{L}_{\text{ideo}} = \sum_{t \in T_{\text{ideo}}} \left[ \left\| \mathbf{t}^{(a)} - \mathbf{t}^{(p)} \right\|_2 - \left\| \mathbf{t}^{(a)} - \mathbf{t}^{(n)} \right\|_2 + \delta_{\text{ideo}} \right] + \]

Story: Trump is tested positive for COVID-19
Continued pretraining objectives

- Story objective: Prevent model from relying on media specific shortcuts
Continued pretraining objectives

- Entity and sentiment focused
- MLM objective: Upsample entity and sentiment tokens
Main results

- Naive MLM training does not help on evaluated tasks
- Proposed objectives help
Further analysis

- **POLITICS** is especially good at
  - Long documents
  - Formal texts
  - Few-shot learning scenarios
Code is available at [https://github.com/launchnlp/POLITICS](https://github.com/launchnlp/POLITICS). Dataset is available upon request. Pretrained POLITICS is available on Huggingface.

Contact: {yujianl,xlfzhang,wangluxy}@umich.edu