



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

Yujian Liu^{1,*}, Xinliang Frederick Zhang^{1,*}, David Wegsman¹, Nick Beauchamp², and Lu Wang¹

¹Computer Science and Engineering, University of Michigan,

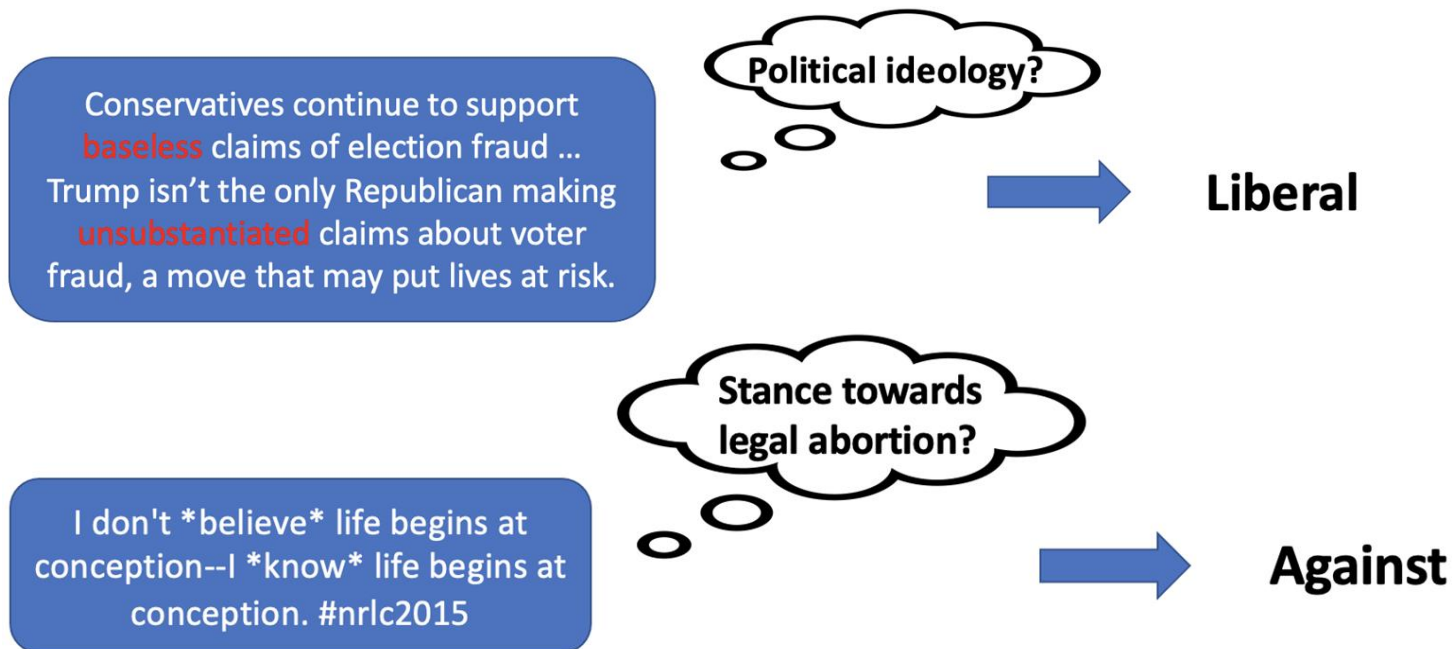
²Political Science and Network Science Institute, Northeastern University

*Equal contributions



NAACL 2022

Analyze political text



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

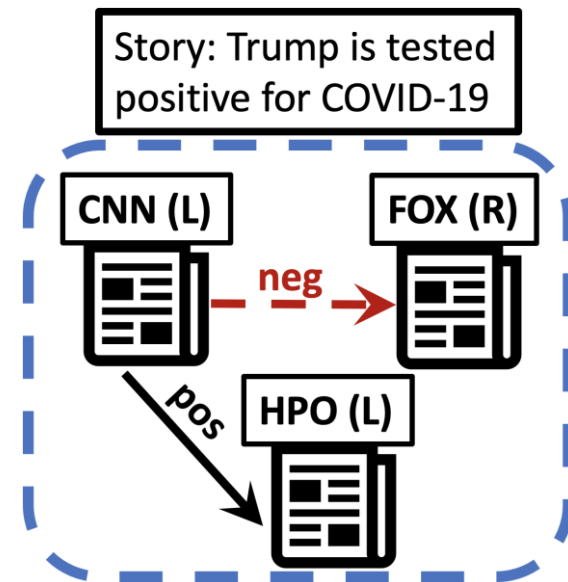
	Daily Kos	HPO	CNN	WaPo	NYT	USA Today	AP	The Hill	TWT	FOX	Breitbart
Ideology	L	L	L	L	L	C	C	C	R	R	R
# articles	100,828	241,417	64,988	198,529	173,737	170,737	279,312	322,145	243,181	330,166	206,512
# words	738.7	729.9	655.7	803.2	599.4	691.7	572.3	426.3	522.7	773.5	483.5

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

Continued pretraining objectives

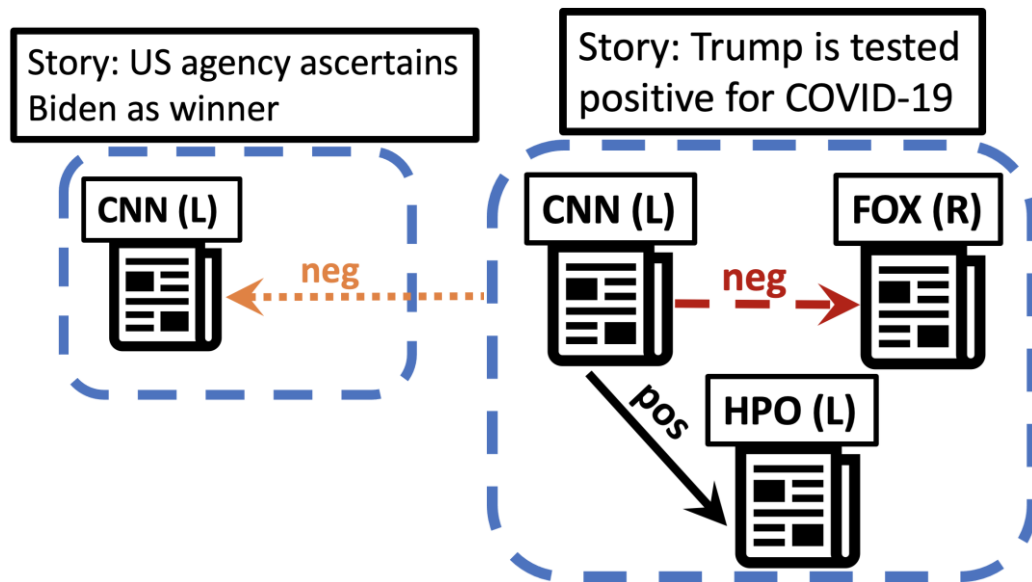
- Ideology objective: Acquire ideology-informed representations

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

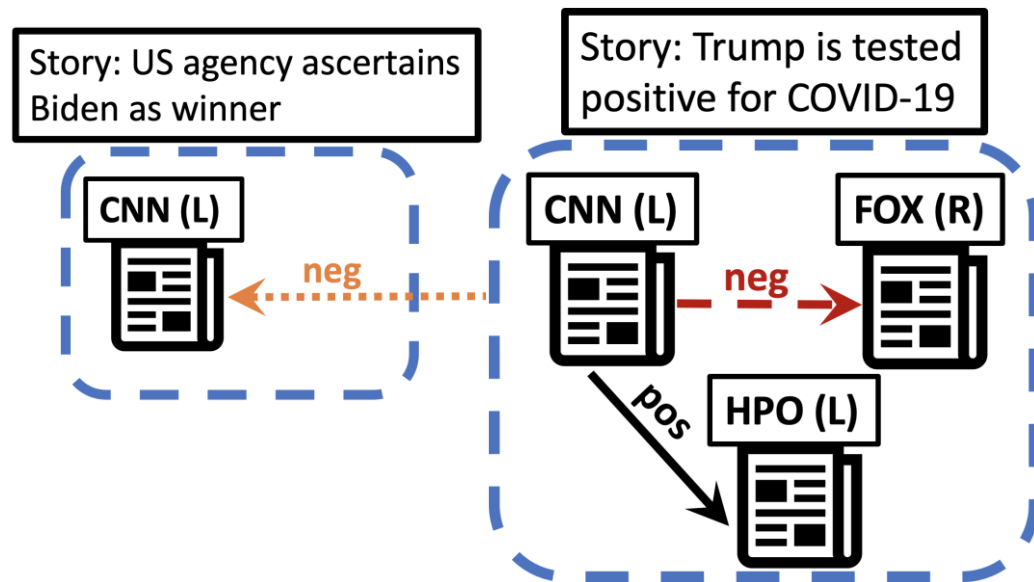


Continued pretraining objectives

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



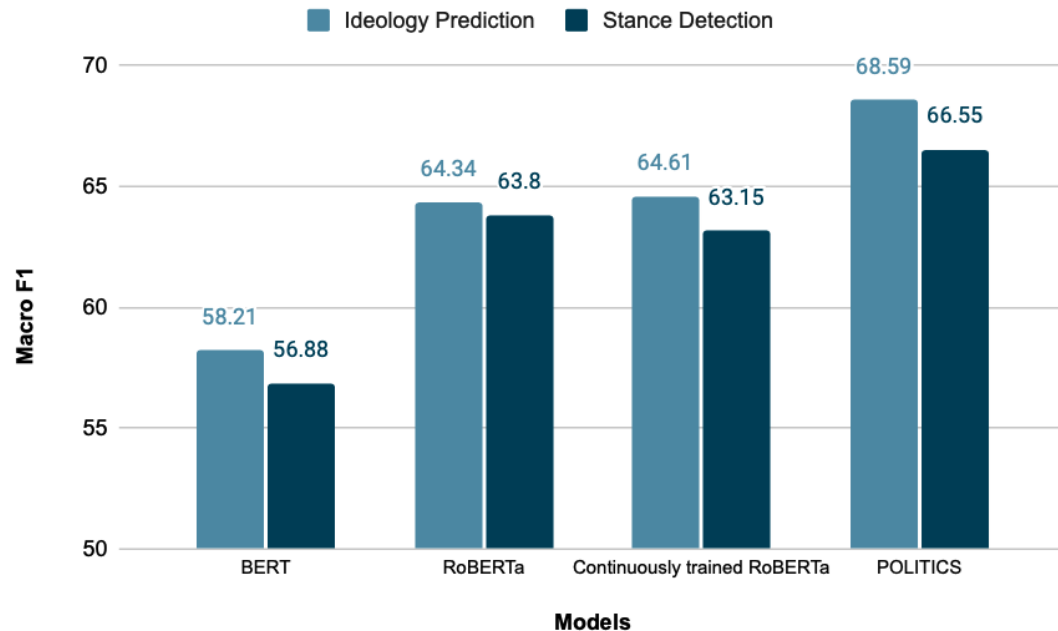
Continued pretraining objectives



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

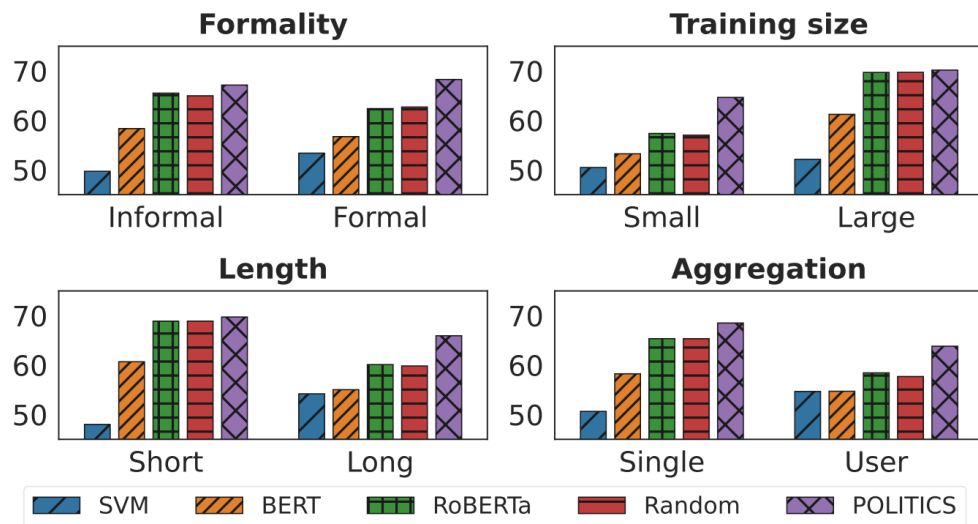
Main results

- 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>.
Dataset is available upon [request](#).
Pretrained *POLITICS* is available on [Huggingface](#).



Research supported by NSF and
UM Advanced Research Computing



Codebase



Dataset



Huggingface

Thanks!