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

Yujian Liu1,∗ Xinliang Frederick Zhang1,2 David Wegsman1 Nick Beauchamp2 Lu Wang1
1Computer Science and Engineering, University of Michigan, Ann Arbor, MI
2Department of Political Science, Northeastern University, Boston, MA
{yujianl, xlfzhang, dwegs, wangluxy}@umich.edu
n.beauchamp@northeastern.edu

Abstract

Ideology is at the core of political science research. Yet, there still does not exist general-purpose tools to characterize and predict ideology across different genres of text. To this end, we study Pretrained Language Models using novel ideology-driven pretraining objectives that rely on the comparison of articles on the same story written by media of different ideologies. We further collect a large-scale dataset, consisting of more than 3.6M political news articles, for pretraining. Our model POLITICS outperforms strong baselines and the previous state-of-the-art models on ideology prediction and stance detection tasks. Further analyses show that POLITICS is especially good at understanding long or formally written texts, and is also robust in few-shot learning scenarios.

1 Introduction

Ideology is an ubiquitous factor in political science, journalism, and media studies (Mullins, 1972; Freedeen, 2006; Martin, 2015). Decades of work has gone into measuring ideology based on voting data (Poole and Rosenthal, 1985; Lewis et al., 2021), survey results (Preoțiuc-Pietro et al., 2017; Ansolabehere et al., 2008; Kim and Fording, 1998; Gabel and Huber, 2000), social networks (Barberá et al., 2015), campaign donation records (Bonica, 2013), and textual data (Laver et al., 2003; Diermeier et al., 2012; Gentzkow et al., 2019; Volkens et al., 2021). Each of those approaches has its strengths and weaknesses. For instance, many political figures do not have voting records; surveys are expensive and politicians are often unwilling to disclose ideology. By contrast, political text is abundant, ubiquitous, yet challenging to work with since language is complex in nature, often domain-specific, and generally unlabeled. There thus remains a strong need for general-purpose tools for measuring ideology using text that can be applied across multiple genres.

∗ Equal contribution by the first two authors.

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 every body 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.

Figure 1: Article snippets by different media on the same news story. Contents that indicate stances and ideological leanings are highlighted in bold (for subjective phrases) and in italics (for objective events).

Using text as data, computational models for ideology measurement have rapidly expanded and diversified, including classical machine learning methods such as ideal point estimation (Groseclose et al., 1999; Shor and McCarty, 2011), Naïve Bayes (Evans et al., 2007), support vector machines (Yu et al., 2008), latent variable models (Barberá et al., 2015), and regression (Peterson and Spirling, 2018); and more recent neural architectures like recurrent neural networks (Iyyer et al., 2014) and Transformers (Baly et al., 2020; Liu et al., 2021). Nonetheless, most of those models leverage datasets with ideology labels drawn from a single domain, and it is unclear if any of them can be generalized to diverse genres of text.

Trained on massive quantities of data, Pretrained Language Models (PLMs) have achieved state-of-the-art performance on many text classification problems, with an additional fine-tuning stage on labeled task-specific samples (Devlin et al., 2019; Liu et al., 2021). Each of those approaches has its strengths and weaknesses. For instance, many political figures do not have voting records; surveys are expensive and politicians are often unwilling to disclose ideology. By contrast, political text is abundant, ubiquitous, yet challenging to work with since language is complex in nature, often domain-specific, and generally unlabeled. There thus remains a strong need for general-purpose tools for measuring ideology using text that can be applied across multiple genres.

∗ Equal contribution by the first two authors.
et al., 2019). Though PLMs suggest the promise of generalizable solutions, their ability to acquire the knowledge needed to detect complex features such as ideology from text across genres remains an open question. PLMs have been shown to capture linguistic structures with a local focus, such as task-specific words, syntactic agreement, and semantic compositionality (Clark et al., 2019; Jawahar et al., 2019). Although word choice is indicative of ideology, ideological leaning and stance are often revealed by which entities and events are selected for presentation (Hackett, 1984; Christie and Martin, 2005; Enke, 2020), with the most notable strand of work in framing theory (Entman, 1993, 2007). One such example is demonstrated in Figure 1, where Daily Kos criticizes Trump’s dishonesty while The Washington Times and Breitbart emphasize the good condition of his health.

In this work, we propose to train PLMs for a wide range of ideology-related downstream tasks. We argue that it is critical for PLMs to consider the global context of a given article. For instance, as pointed out by Fan et al. (2019), one way to acquire such context is through comparison of news articles on the same story but reported by media of different ideologies. Given the lack of suitable datasets, we first collect a new large-scale dataset, BIGNEWS. It contains 3,689,229 English news articles on politics, gathered from 11 United States (US) media outlets covering a broad ideological spectrum. We further downsample and cluster articles in BIGNEWS by different media into groups, each consisting of pieces aligned on the same story. The resultant dataset, BIGNEWSALIGN, contains 1,060,512 stories with aligned articles.

Next we train a new PLM, POLITICS, based on a Pretraining Objective Leveraging Inter-article Triplet-loss using Ideological Content and Story. Concretely, we leverage continued pretraining (Gururangan et al., 2020), where we design an ideology objective operating over clusters of same-story articles to compact articles with similar ideology and contrast them with articles of different ideology. The learned representation can better discern the embedded ideological content. We further enhance it with a story objective that ensures the model to focus on meaningful content instead of overly relying on shortcuts, e.g., media boilerplate. Both objectives are used together with our specialized masked language model objective that focuses on entities and sentiments to train POLITICS.

Our main goal here is to create general-purpose tools for analyzing ideological content for researchers and practitioners in the broad community. Furthermore, when experimenting on 11 ideology prediction and stance detection tasks using 8 datasets of different genres, including a newly collected dataset from AllSides, POLITICS outperforms both a strong SVM baseline and previous PLMs on 8 tasks. Notably, POLITICS is particularly effective on long documents, e.g., achieving 10% improvements on both ideology prediction and stance detection tasks over RoBERTa (Liu et al., 2019). We further show that our model is more robust in setups with smaller training sets.

2 Related Work

**Ideology prediction** is a critical task for quantitative political science (Mullins, 1972; Freeden, 2006; Martin, 2015; Wilkerson and Casas, 2017). Both classical methods (e.g., Naive Bayes, SVM; Evans et al., 2007; Yu et al., 2008; Sapiro-Gheiler, 2019) and deep learning models (e.g., RNN; Iyyer et al., 2014) have been used to predict ideology on a variety of datasets where ideology labels are available, such as legislative speeches (Laver et al., 2003) and U.S. Supreme Court briefs (Evans et al., 2007). Notably, Liu et al. (2021) pretrains a Transformer-based language generator to minimize the ideological bias in generated text. As generative models are not as effective as masked language models (MLMs) at text classification, our goal differs in that we train MLMs to recognize ideological contents in various domains and tasks.

**Stance detection** is a useful task for ideology analysis because co-partisans are generally positive towards each other and negative towards counter-partisans (Aref and Neal, 2021). There has been a large body of work on identifying individuals’ stances towards specific targets from the given text (Thomas et al., 2006; Walker et al., 2012; Hasan and Ng, 2013). On the methodology side, Mohammad et al. (2016b) and Küçük and Can (2018) apply statistical models, e.g., SVM, with handcrafted text features. Neural methods have also been widely investigated, including CNN (Wei et al., 2016), LSTM (Augenstein et al., 2016), hierarchical networks (Sun et al., 2018), and representation learning (Darwish et al., 2020).

Recent research focus resides in leveraging
PLMs for predicting stances, e.g., incorporating extra features (Prakash and Madabushi, 2020). Kawintiranon and Singh (2021) share a similar spirit with our work by upsampling tokens to mask. However, they pre-define a list of tokens customized for the given targets, which is hard to generalize to new targets. We aim to train PLMs relying on general-purpose sentiment lexicons and important entities, to foster model generalizability.

**Domain-specific Pretrained Language Models.**

PLMs, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have obtained state-of-the-art results on many NLP tasks. Inspired by the observation that a continued pretraining phase on in-domain data yields better performance (Gururangan et al., 2020), domain-specific PLMs are introduced (Beltagy et al., 2019; Yang et al., 2020; Huang et al., 2019; Lee et al., 2020). However, they only use the default MLM objective, without considering domain knowledge. In this work, we design ideology-driven pretraining objectives to inject domain knowledge to discern ideologies and related stances.

Focusing on the news domain, PLMs have been primarily used for factuality prediction (Jwa et al., 2019; Zellers et al., 2019; Kaliyar et al., 2021) and topic classification (Liu et al., 2020; Büyükköz et al., 2020; Gupta et al., 2020) by fine-tuning on task-specific datasets. Few work has investigated PLMs for understanding political ideology evinced in texts. One exception is Baly et al. (2020), where they also leverage the triplet loss as the pretraining objective. However, our work is novel in at least three aspects. First, our triplet loss is designed to capture the ideological (dis)similarity among articles on the same story, while the loss used by Baly et al. (2020) operates on articles of the same topic. As a result, their approach can falsely compact representations of very different news contents, e.g., articles on “Japan Economics” and “Indian Troops” both belong to the topic of “Asia”. Moreover, our newly introduced story objective can effectively prevent the model from relying on media-specific language (e.g., “for the New York Times”), while their objective may fail to do so, and thus lacks generalizability to languages used by different media and other ideology-related tasks. Finally, we use BIGNES that contains more than 3M articles, which is more suitable for pretraining large models than the small dataset (35k articles) used by Baly et al. (2020). To the best of our knowledge, we are the first to systematically study and release PLMs for ideology-related study in the US political domain.

### 3 Pretraining Datasets

#### 3.1 Data Crawling

We collect pretraining datasets from online news articles with diverse ideological leanings and language usage. We select 11 media outlets based on their ideologies (far-left to far-right) and popularity.\(^2\) We convert their ideologies into three categories: left, center, and right, and crawl all pages published by them between January 2000 and June 2021, from Common Crawl and Internet Archive. We then follow Raffel et al. (2020) for data cleaning, and, additionally, only retain news articles related to US politics. Appendix A describes in detail the steps for removing non-articles pages, duplicates, non-US pages, and boilerplate.

The cleaned data, dubbed BIGNES, contains 3,689,229 US political news articles. To mitigate the bias that some media dominate the model training, we downsample the corpus so that each ideology contributes equally. The downsampled corpus, BIGNESBLN, contains 2,331,552 news articles, with statistics listed in Table 1. We keep 30K held-out articles as validation set.

---

3.2 Aligning Articles on the Same Story

We compare how media outlets from different sides report the same story, which intuitively better captures ideological content. To this end, we design an algorithm to align articles in BIGNEWSBLN that cover the same story. We treat each article as an anchor, and find matches from other outlets based on the following similarity score:

\[
\text{sim}(p_i, p_j) = \alpha \cdot \text{sim}_t(p_i, p_j) + (1 - \alpha) \cdot \text{sim}_e(p_i, p_j) \tag{1}
\]

where \(p_i\) and \(p_j\) are two articles, \(\text{sim}_t\) is the cosine similarity between TF-IDF vectors of \(p_i\) and \(p_j\), \(\text{sim}_e\) is the weighted Jaccard similarity between the sets of named entities\(^3\) in \(p_i\) and \(p_j\), and \(\alpha = 0.4\) is a hyperparameter. During alignment, for an article from an outlet to be considered as a match, it must be published within three days before or after the anchor, has the highest similarity score among articles from the same outlet, and the score is at least \(\theta = 0.23\). Hyperparameters \(\alpha\) and \(\theta\) are searched on the Basil dataset (Fan et al., 2019), which contains manually aligned articles.\(^4\) After deduplicating articles in each story cluster, we obtain BIGNEWSALIGN, containing 1,060,512 clusters with an average of 4.29 articles in each. Appendix B details the alignment algorithm.

4 POLITICS via Continued Pretraining

Here we introduce our continued pretraining methods based on a newly proposed ideology objective that drives representation learning to better discern ideological content by comparing same-story articles written by media of different ideologies. As shown in Figure 2, given a story cluster, we choose an article published by media on the left or right as the anchor. We then take articles in the same cluster with the same ideology as positive samples, and articles with the opposite ideology as negative ones. The ideology objective is formulated as follows:

\[
\mathcal{L}_{\text{ideo}} = \sum_{t \in \mathcal{T}_{\text{ideo}}} \left[ \left\| t^{(a)} - t^{(p)} \right\|_2 - \left\| t^{(a)} - t^{(n)} \right\|_2 + \delta_{\text{ideo}} \right]_+	ag{2}
\]

where \(\mathcal{L}_{\text{ideo}}\) is the set of all ideology triplets, \(t^{(a)}\), \(t^{(p)}\), and \(t^{(n)}\) are the [CLS] representations of anchor, positive, and negative articles in triplet \(t\), \(\delta_{\text{ideo}}\) is a hyperparameter, and \(\lceil \cdot \rceil_+\) is \(\max(\cdot, 0)\).

Next, we augment the ideology objective with a story objective to allow the model to focus on semantically meaningful content and to prevent the model from focusing on “shortcuts” (such as media-specific languages) to detect ideology. To construct story triplets, we use the same <anchor, positive> pairs as in the ideology triplet, and then treat articles from the same media outlet but on different stories as negative samples, as depicted in Figure 2. Similarly, our story objective is formulated as follows:

\[
\mathcal{L}_{\text{story}} = \sum_{t \in \mathcal{T}_{\text{story}}} \left[ \left\| t^{(a)} - t^{(p)} \right\|_2 - \left\| t^{(a)} - t^{(n)} \right\|_2 + \delta_{\text{story}} \right]_+	ag{3}
\]

where \(\mathcal{T}_{\text{story}}\) contains all story triplets, and \(\delta_{\text{story}}\) is a hyperparameter searched on the validation set.

\(^3\) Extracted by Stanford CoreNLP (Manning et al., 2014).
\(^4\) Our algorithm achieves a mean reciprocal rank of 0.612 on Basil, with detailed evaluation in Appendix B.
4.2 Entity- and Sentiment-aware MLM

Here we present a specialized MLM objective to collaborate with our triplet loss based objectives for better representation learning. Notably, political framing effect is often reflected in which entities are selected for reporting (Gentzkow et al., 2019). Moreover, the occurrence of sentimental content along with the entities also signal stances (Mohammad et al., 2016b). Therefore, we take a masking strategy that upsamples entity tokens (Sun et al., 2019; Guu et al., 2020; Kawintiranon and Singh, 2021) and sentiment words to be masked for the MLM objective, which improves from prior pre-training work that only considers article-level comparison (Baly et al., 2020).

Concretely, we consider named entities with types of PERSON, NORP, ORG, GPE and EVENT. We detect sentiment words using lexicons by Hu and Liu (2004) and Wilson et al. (2005). To allow MLM training to focus on entities and sentiment, we mask them with a 30% probability, and then randomly mask remaining tokens until 15% of all tokens are reached, as done in Devlin et al. (2019). Masked tokens are replaced with [MASK], random tokens, and original tokens with a ratio of 8:1:1.

4.3 Overall Pretraining Objective

We combine the aforementioned objectives as our final pretraining objective as follows:

\[ L = \beta L_{\text{ideology}} + \gamma L_{\text{story}} + (1 - \beta - \gamma) L_{\text{MLM}} \] (4)

where \( \beta = \gamma = 0.25 \). Using \( L \), POLITICS is produced via continued training on RoBERTa (Liu et al., 2019).\(^5\) We do not try to train the model from scratch since BigNEWSBLN only has \( \sim 10\)GB data, smaller than corpus for RoBERTa (~160GB). Hyperparameters are listed in Table A5.

5 Experiments

Given the importance of ideology prediction and stance detection tasks in political science (Thomas et al., 2006; Wilkerson and Casas, 2017; Chatsiou and Mikhaylov, 2020), we conduct extensive experiments on a wide spectrum of datasets with 11 tasks (§5.1). We then compare with both classical models and prior PLMs (§5.2), and among our model variants (§5.3). We present and discuss results in §5.5, where POLITICS outperform all three baselines on 8 out of 11 tasks. For all models, MLM objectives are trained with BigNEWSBLN, and ideology and story objectives are trained on BigNEWSALIGN. Details are in Appendix C.1.

5.1 Datasets and Tasks

Our tasks are discussed below, with statistics listed in Table 2 and more descriptions in Appendix D.

**Ideology prediction** tasks for predicting the political leanings are evaluated on the following datasets.

- Congress Speech (CongS; Gentzkow et al., 2018) contains speeches from US congressional records, each labeled as liberal or conservative.
- AllSides (AllS; new) is a website that assesses political bias and ideology of US media. In this study, we collect articles from AllSides with their ideological leanings on a 5-point scale.
- Hyperpartisan (HP; Kiesel et al., 2019) is a shared task of predicting a binary label for an article as being hyperpartisan or not. We convert it into a 3-way classification task by splitting hyperpartisan news into left and right.
- YouTube (YT; Wu and Resnick, 2021) contains discussions on YouTube. cmt. and user refer to predicting left/right at the comment- and user-level, respectively.
- Twitter (TW; Preotiuc-Pietro et al., 2017) collects a group of Twitter users with self-reported ideologies on a 7-point scale. We merge them into 3-way labels.

**Stance detection** tasks, which predict a subject’s attitude towards a given target from a piece of text, are listed below. All tasks take a 3-way label

---

\(^5\) We use roberta-base model card from Huggingface.

Table 3: Macro F1 scores on 11 evaluation tasks (average of 5 runs). Tasks are sorted by text length, short to long, within each group. “All avg” is the average of all 11 tasks. Best results are in bold and second best are underlined. Our models with triplet-loss objectives that outperform RoBERTa are in blue. Our models with specialized sampling methods that outperform the one with vanilla MLM (Random) are in green. POLITICS uses Ideology + Story Obj. and Upsamp. Ent. + Sentiment. Results where POLITICS outperforms all baselines are in red, with ∗ indicating statistical significance (Mann–Whitney U test; Mann and Whitney, 1947, \( p \leq 0.05 \)). Standard deviations (std) are reported in Table A11. The range of std over tasks is [0.31, 3.42] for POLITICS, and [0.48, 7.35] for RoBERTa.

<table>
<thead>
<tr>
<th>Model</th>
<th>YT (cmnt.)</th>
<th>ConS</th>
<th>HP</th>
<th>AllS</th>
<th>YT (user)</th>
<th>TW</th>
<th>Ideo. avg</th>
<th>SIn (seen)</th>
<th>SIn (unseen)</th>
<th>Basline (sent.)</th>
<th>VAST</th>
<th>Pol. (art.)</th>
<th>Pol. (seen)</th>
<th>All avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>65.34</td>
<td>71.31</td>
<td>61.25</td>
<td>52.51</td>
<td>66.89</td>
<td>82.85</td>
<td>59.96</td>
<td>51.18</td>
<td>32.89</td>
<td>51.08</td>
<td>39.54</td>
<td>30.77</td>
<td>41.09</td>
<td>51.38</td>
</tr>
<tr>
<td>BERT</td>
<td>64.64</td>
<td>65.88</td>
<td>48.42</td>
<td>60.88</td>
<td>65.24</td>
<td>44.20</td>
<td>58.21</td>
<td>40.39</td>
<td>62.81</td>
<td>70.53</td>
<td>45.63</td>
<td>56.88</td>
<td>57.61</td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>66.72</td>
<td>67.25</td>
<td>60.43</td>
<td>74.75</td>
<td>67.98</td>
<td>48.90</td>
<td>64.34</td>
<td>70.15</td>
<td>63.08</td>
<td>68.16</td>
<td>76.25</td>
<td>41.36</td>
<td>63.80</td>
<td>64.09</td>
</tr>
<tr>
<td>Baly et al. (2020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with Original Data</td>
<td>65.42</td>
<td>66.74</td>
<td>58.37</td>
<td>72.89</td>
<td>70.47</td>
<td>44.95</td>
<td>63.14</td>
<td>68.66</td>
<td>56.29</td>
<td>61.30</td>
<td>75.57</td>
<td>37.98</td>
<td>59.96</td>
<td>61.69</td>
</tr>
<tr>
<td>with BizNewsBLN</td>
<td>68.57</td>
<td>70.39</td>
<td>71.24</td>
<td>76.47</td>
<td>74.74</td>
<td>47.38</td>
<td>68.13</td>
<td>65.84</td>
<td>69.54</td>
<td>64.60</td>
<td>75.03</td>
<td>41.84</td>
<td>58.57</td>
<td>63.79</td>
</tr>
<tr>
<td>MLP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our models with triplet loss objective only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ideology Obj.</td>
<td>66.20</td>
<td>68.18</td>
<td>64.15</td>
<td>76.52</td>
<td>68.15</td>
<td>42.66</td>
<td>64.31</td>
<td>68.78</td>
<td>59.61</td>
<td>64.18</td>
<td>76.03</td>
<td>44.94</td>
<td>62.71</td>
<td>63.58</td>
</tr>
<tr>
<td>Story Obj.</td>
<td>66.09</td>
<td>69.11</td>
<td>56.70</td>
<td>74.59</td>
<td>68.89</td>
<td>46.53</td>
<td>63.65</td>
<td>69.02</td>
<td>63.54</td>
<td>67.21</td>
<td>76.66</td>
<td>53.16</td>
<td>65.92</td>
<td>64.68</td>
</tr>
<tr>
<td>Ideology Obj. + Story Obj.</td>
<td>68.91</td>
<td>69.10</td>
<td>63.08</td>
<td>76.23</td>
<td>77.58</td>
<td>48.98</td>
<td>67.31</td>
<td>66.63</td>
<td>63.17</td>
<td>64.37</td>
<td>76.18</td>
<td>47.01</td>
<td>64.08</td>
<td>65.84</td>
</tr>
<tr>
<td>Our models with masked language model objective only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>67.82</td>
<td>70.32</td>
<td>60.59</td>
<td>73.54</td>
<td>70.77</td>
<td>44.62</td>
<td>64.61</td>
<td>69.16</td>
<td>60.39</td>
<td>69.94</td>
<td>77.11</td>
<td>39.16</td>
<td>63.15</td>
<td>63.95</td>
</tr>
<tr>
<td>Upsamp. Ent.</td>
<td>69.06</td>
<td>70.32</td>
<td>60.99</td>
<td>70.99</td>
<td>71.40</td>
<td>47.16</td>
<td>64.82</td>
<td>69.81</td>
<td>63.08</td>
<td>69.49</td>
<td>76.76</td>
<td>46.46</td>
<td>65.12</td>
<td>64.96</td>
</tr>
<tr>
<td>Upsamp. Sentiment</td>
<td>67.41</td>
<td>70.03</td>
<td>56.05</td>
<td>72.35</td>
<td>74.93</td>
<td>48.15</td>
<td>64.82</td>
<td>70.09</td>
<td>60.81</td>
<td>71.28</td>
<td>76.61</td>
<td>44.42</td>
<td>64.64</td>
<td>64.74</td>
</tr>
<tr>
<td>Upsamp. Ent. + Sentiment</td>
<td>68.31</td>
<td>71.42</td>
<td>58.02</td>
<td>71.90</td>
<td>71.04</td>
<td>47.31</td>
<td>64.67</td>
<td>69.25</td>
<td>62.84</td>
<td>69.23</td>
<td>77.10</td>
<td>43.16</td>
<td>64.32</td>
<td>64.51</td>
</tr>
<tr>
<td>POLITICS</td>
<td>67.83</td>
<td>70.86</td>
<td>50.25</td>
<td>74.93</td>
<td>76.75</td>
<td>48.92</td>
<td>68.39</td>
<td>69.41</td>
<td>61.26</td>
<td>75.41</td>
<td>76.75</td>
<td>51.94</td>
<td>66.58</td>
<td>67.86</td>
</tr>
</tbody>
</table>

5.2 Baselines

We consider three baselines. First, we train a linear SVM using unigram and bigram features for each task, since it is a common baseline in political science (Yu et al., 2008; Diermeier et al., 2012). Hyperparameters and feature selection are described in Table A8. We further compare with BERT and RoBERTa, following the standard fine-tuning process for ideology prediction tasks and using the prompt described in §5.4 for stance detection.

5.3 Model Variants

We consider several variants of POLITICS. First, using triplet loss objective only, we experiment on models trained with ideology objective (Ideology Obj.), story objective (Story Obj.), or both.

Next, we continue pretaining RoBERTa with MLM objective only, using vanilla MLM objective (Random), entity focused objective (Upsamp. Ent.), sentiment focused objective (Upsamp. Sentiment), or upsampling both entity and sentiment.

5.4 Fine-tuning Procedure

We fine-tune each neural model for up to 10 epochs, with early stopping enabled. We select the best fine-tuned model on validation sets using F1. Details of experimental setups are in Table A7.

Ideology Prediction. We follow common practice of using the [CLS] token for standard fine-tuning (Devlin et al., 2019). For Twitter and YouTube User data, we encode them using sliding windows and aggregate by mean pooling.

Stance Detection. We follow Schick and Schütze (2021) on using prompts to fine-tune models for stance detection. We curate 11 prompts (in Table A8) and choose the best one based on the average F1 by RoBERTa on all stance detection tasks: \( p[\text{SEP}] \) The stance towards [target] is [MASK]. The model is trained to predict [MASK] for stance, conditioned on the input \( p \) and [target].
5.5 Main Results

Table 3 presents F1 scores on all tasks. POLITICS achieves the best overall average F1 score across the board, 3.6% better than the strongest baseline, RoBERTa. More importantly, POLITICS alone outperforms all three baselines listed in §5.2 on 8 out of 11 tasks, including more than 10% of improvement for ideology labeling on Hyperpartisan and Youtube user-level. We attribute the performance gain to our proposed ideology-driven pretraining objective, which helps capture partisan content. Note that, on some tasks, other model variants lead POLITICS by a small margin, and this may be of interest to practitioners performing specific tasks.

We further compare with the model proposed by Baly et al. (2020), which also leverages triplet loss as pretraining objective but on articles of the same topics. We implement two versions of their model, using the original data released by Baly et al. (2020)\(^7\) and our BigNewsBLN. First, pretraining on our BigNewsBLN yields better results on ideology prediction tasks than using the original data, indicating the value of BigNewsBLN. Second, using the triple construction method by Baly et al. (2020) with BigNewsBLN does not generalize well on the stance detection task, compared to POLITICS and its variants. This highlights the advantage of our objectives that enable content comparison among articles of the same stories.

Moreover, our ideology-driven objectives help acquire knowledge needed to discern ideology as well as stance detection. When equipping the RoBERTa model with ideology and story objectives but no MLM objective, it achieves the second best overall performance.

Next, focusing on entities better identifies stance. Simply continuing training RoBERTa with vanilla MLM objective (Random) does not yield performance gain on stance detection, while our upsampling methods make a difference, i.e., increasing sampling ratios of entities improves F1 by 2%.

Comparisons with Previous State-of-the-arts.

Using the original binary prediction setup (i.e., hyperpartisan or not) on Hyperpartisan data (Kiesel et al., 2019), POLITICS obtains an accuracy of 85.2, leading previous state-of-the-art results by at least 3 points, as shown in Table 4.

POLITICS achieves an F1 of 77.0 on the original VAST data where the previous state-of-the-art model obtained 69.2 (Jayaram and Allaway, 2021).

On SemEval, POLITICS yields an F1 of 71.3 where the best performance is 76.5 by Al-Ghadir et al. (2021). Notably, we adopt one single classifier in our setup, while they include separate models for different prediction targets, which have been shown to outperform one single classifier (Mohammad et al., 2016a). Full comparisons with previous methods are included in Appendix F. For other tasks, there is no direct comparison as the datasets are either originally used for different prediction tasks (e.g., Basil is used for detecting media bias spans) or newly collected by this work.

On Texts of Different Characteristics.

Based on Table 2, we further study the model’s performance on data of different properties: language formality, training size, document length, and aggregation level. As shown in Figure 3, with each property (concrete criterion in Appendix E), we divide tasks into two categories. POLITICS yields greater improvements on more formal and longer text, since pretraining is done on news articles. POLITICS is also more robust to training sets with

---

\(^7\)https://github.com/ramybaly/Article-Bias-Prediction.

<table>
<thead>
<tr>
<th>Models</th>
<th>Hyperpartisan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc.</td>
<td>Precision</td>
</tr>
<tr>
<td>Jiang et al., 2019 (ELMo+CNN)</td>
<td>82.2</td>
</tr>
<tr>
<td>Srivastava et al., 2019 (Logistic Regression)</td>
<td>82.0</td>
</tr>
<tr>
<td>Hanawa et al., 2019 (BERT)</td>
<td>80.9</td>
</tr>
<tr>
<td>Ishibash and Johansson, 2019 (SVM)</td>
<td>80.6</td>
</tr>
<tr>
<td>Yeh et al., 2019 (ULMFiT)</td>
<td>80.3</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>84.3</td>
</tr>
<tr>
<td>POLITICS</td>
<td>85.2</td>
</tr>
</tbody>
</table>

Table 4: Comparison with the previous state-of-the-art models on Hyperpartisan using the original binary prediction setup. Best results are in bold. POLITICS obtains the best accuracy, recall and F1. Note, precision, recall, and F1 are measured for the hyperpartisan class.
Table 5: Ablation study results on POLITICS. POLITICS + Ideo. Pred.: triplet-loss objective is replaced with a hard label prediction objective on ideology of articles (left vs. right). Best results are in bold. Darker red shows greater improvements. Darker blue indicates larger performance drop. The ideology objective contributes the most to POLITICS, followed by the story objective.

<table>
<thead>
<tr>
<th></th>
<th>VT</th>
<th>Cong</th>
<th>HP</th>
<th>AIB</th>
<th>VT</th>
<th>TW</th>
<th>Ideo.</th>
<th>SEval</th>
<th>SEval</th>
<th>Basil</th>
<th>VAST</th>
<th>Basil</th>
<th>Stan.</th>
<th>All avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Ideology Obj.</td>
<td>-3.78</td>
<td>-2.17</td>
<td>-16.35</td>
<td>-3.28</td>
<td>-12.54</td>
<td>-3.43</td>
<td>-6.93</td>
<td>-0.38</td>
<td>-0.83</td>
<td>-2.22</td>
<td>-0.45</td>
<td>-16.01</td>
<td>-4.38</td>
<td>-5.77</td>
</tr>
<tr>
<td>No Story Obj.</td>
<td>+1.98</td>
<td>+0.64</td>
<td>-0.72</td>
<td>+0.70</td>
<td>+0.29</td>
<td>-1.78</td>
<td>+0.19</td>
<td>-1.23</td>
<td>+2.94</td>
<td>+3.36</td>
<td>-0.87</td>
<td>-10.75</td>
<td>-2.66</td>
<td>-1.11</td>
</tr>
<tr>
<td>No Upsamp. Ext.</td>
<td>+0.18</td>
<td>-0.05</td>
<td>-0.05</td>
<td>+0.55</td>
<td>-0.29</td>
<td>-1.20</td>
<td>-0.24</td>
<td>+0.62</td>
<td>-0.67</td>
<td>-3.74</td>
<td>-0.55</td>
<td>-1.20</td>
<td>-1.11</td>
<td>-0.64</td>
</tr>
<tr>
<td>No Upsamp. Sentiment</td>
<td>+0.75</td>
<td>-0.28</td>
<td>+0.22</td>
<td>-1.27</td>
<td>-0.11</td>
<td>-1.40</td>
<td>-0.35</td>
<td>-0.84</td>
<td>-1.67</td>
<td>-3.91</td>
<td>-1.10</td>
<td>+1.44</td>
<td>-0.55</td>
<td>-0.44</td>
</tr>
<tr>
<td>POLITICS + Ideo. Pred.</td>
<td>+1.46</td>
<td>+1.10</td>
<td>-1.01</td>
<td>+4.72</td>
<td>+2.02</td>
<td>-3.96</td>
<td>+0.72</td>
<td>+0.41</td>
<td>-0.52</td>
<td>-3.82</td>
<td>+0.12</td>
<td>-3.10</td>
<td>-1.38</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Figure 4: Average of ideology prediction and stance detection performances with few-shot learning. POLITICS uniformly outperforms RoBERTa and it being continued pretrained with vanilla MLM (Random).

small sizes, showing the potential effectiveness in few-shot learning, which is echoed in §6.1.

6 Further Analyses

6.1 Few-shot Learning

We first fine-tune all PLMs on small numbers of samples. POLITICS consistently outperforms the two counterparts on both tasks, using small training sets (Figure 4). More importantly, naively training RoBERTa on the large BigNewsBLN does not help ideology prediction. By contrast, our ideology-driven objective can better capture ideology than the baselines, even when using only 16 samples for fine-tuning on the ideology tasks.

6.2 Ablation Study on POLITICS

We show the impact of removing each ideology-driven pretraining objective and upsampling strategy from POLITICS in Table 5. First, removing the ideology objective results in the most loss on both tasks. This again demonstrates the effectiveness of our triplet-loss formulation over same-story articles. Removing the story objective also hurts the overall performance by 1% but improves the ideology prediction marginally. This shows that the story objective functions as an auxiliary constraint to avoid over-fitting on the “shortcuts” for discerning ideologies. Moreover, removing upsampling strategies generally weakens POLITICS’s performance, but only to a limited extent.

We also experiment with a setup with hard-ideology learning (i.e., directly predicting the ideology of each article without using triplet-loss objectives). Not surprisingly, this variant (POLITICS + Ideo. Pred.) outperforms POLITICS on ideology prediction since it can directly learn ideology from the annotated labels. However, it has been overfitted to ideology prediction tasks and lacks generalizability, thus yields worse performance on stance detection.

6.3 Visualizing Attentions

On the Hyperpartisan task, we visualize the last layer’s attention weights between the [CLS] token and all other tokens by POLITICS and RoBERTa pretrained with vanilla MLM on BigNewsBLN (Random). We randomly sample 20 test articles, and for 13 of them, POLITICS is able to capture salient entities, events, and sentiments in the text whereas Random cannot. We present one example in Figure 5 where POLITICS captures “Ashley Judd”, “the worst”, and “Trump”. More examples are given in Appendix G. This finding confirms that our ideology-driven objective and upsampling strategies can help the model focus more on entities of political interest as well as better recognize sentiments.
Figure 5: Last layer attention scores between [CLS] token and other input tokens (aggregated over all heads). POLITICS captures “Ashley Judd”, “worst”, and “Trump”.

Figure 6: Model perplexities on 30K validation articles in BigNEWSBLN. Perplexities do not drop sharply on POLITICS compared with RoBERTa being continued pretrained with MLM objective (Random), suggesting POLITICS can yield superior predictive performance while not overfitting with ideological languages.

6.4 POLITICS on Different Ideologies

Finally, we measure whether PLMs would acquire ideological bias as measured by whether they fit with languages used by a specific ideology. Concretely, we follow Salazar et al. (2020) to evaluate PLMs on 30K held-out articles of different ideologies from BigNEWSBLN with pseudo-perplexity. For efficiency, we estimate the pseudo log-likelihood based on 200 random tokens in each article as used by Wang and Cho (2019). As illustrated in Figure 6, while MLM objective (Random) is effective at fitting a corpus, i.e., having the lowest perplexities, triplet-loss objectives act as regularizers during pretraining, shown by the higher perplexity of POLITICS compared to Random. Interestingly, we find center and right articles have lower perplexity than that of left articles. We hypothesize that it relates to political science findings that, over recent periods of political polarization in US, Republicans have become somewhat more coherent and similar than Democrats (Grossmann and Hopkins, 2016; Benkler et al., 2018), and are thus easier to predict.

7 Conclusion

We study the problem of training general-purpose tools for ideology content understanding and prediction. We present POLITICS, trained with novel ideology-driven pretraining objectives based on the comparisons of same-story articles written by media outlets of different ideologies. To facilitate model training, we also collect a large-scale dataset, BigNEWS, consisting of news articles of different ideological leanings. Experiments on diverse datasets for ideology prediction and stance detection tasks show that POLITICS outperforms strong baselines, even with a limited amount of labeled samples for training, and state-of-the-art models.

Acknowledgments

This work is supported in part through National Science Foundation under grants IIS-2100885 and IIS-2127747, and computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. We appreciate the anonymous reviewers for their helpful comments. We thank the members of the LAUNCH group at the University of Michigan for discussions and suggestions. We also thank Changyuan Qiu for helping collect the AllSides articles, Siqi Wu for giving us access to the Youtube comments, and Daniel Preotiuc-Pietro for sharing the Twitter user accounts used in the original study to allow the collection of the corresponding users’ tweets in this work.

8 Ethical Considerations

8.1 BigNEWS Collection

All news articles were collected in a manner consistent with the terms of use of the original sources
as well as the intellectual property and the privacy rights of the original authors of the texts, i.e., source owners. During data collection, the authors honored privacy rights of content creators, thus did not collect any sensitive information that can reveal their identities. All participants involved in the process have completed human subjects research training at their affiliated institutions. We also consulted Section 107\(^8\) of the U.S. Copyright Act and ensured that our collection action fell under the fair use category.

8.2 Dataset Usage

All of the newly collected datasets in this work will be made available upon request. Pretraining corpus details are included in Section 3. The other seven datasets used for downstream evaluation are obtained in the following ways. CongS, HP, BASIL, VAST and SEval are acquired by direct download. CongS is released under the ODC-BY 1.0 license (free to share, create, and adapt). HP and SEval are developed in shared tasks by the NLP community, which allow the use of copyrighted material without permission from the copyright holder for research purposes (Escartín et al., 2017). For VAST, the author explicitly states “We make our dataset and models available for use”. BASIL is developed by the last author and her collaborators. For YT and TW, we consult with the corresponding authors and obtain the datasets by agreeing that we will not further distribute them. Dataset details are listed in Section 5.1 and Appendix D.

8.3 Benefit and Potential Misuse

**Intended use.** The models developed in this work can assist the general public to measure and understand ideological language used in diverse genres of texts. For example, POLITICS can help the general public know where their representatives stand on key issues. Our experiments in Section 5 demonstrate how POLITICS would be deployed in real life when handling applications in both ideology prediction and stance detection. We deem that our extensive experiments have covered the major usage of POLITICS.

**Failure mode** is defined as situations where POLITICS fails to correctly predict the ideology of an individual or a given text. In such cases, POLITICS might deliver misinformation or cause misunderstanding towards a political figure or a policy. For vulnerable populations (e.g., people who maybe not be able to make the right judgements), the harm could be tremendously magnified when they fail to interpret the model outputs or blindly trust machine responses. Ideally, the interpretation of our model’s predictions should be carried out within the broader context of the source text.

**Misuse potential.** Users may mistakenly take the machine prediction as a golden rule or a fact. We would recommend any politics-related machine learning models, including ours, put up an “use with caution” message to encourage users to check more sources or consult political science experts to reduce the risk of being misled by single source. Moreover, POLITICS might also be misused to label people with a specific political leaning that they do not want to be associated with. We suggest that when in use the tools should be accompanied with descriptions about their limitations and imperfect performance, as well as allow users to opt out from being the subjects of measurement.

**Potential limitation.** Although multiple genres are considered, the genre coverage is not exhaustive, and does not include other trending media or content of different modalities for expressing opinions, such as TV transcripts, images, and videos. Thus, the predictive performance of POLITICS may still be under investigated. Further, in downstream evaluation, POLITICS is only trained and tested in the same domain, so its cross-genre ability needs further evaluation.

**Bias Mitigation.** During data preprocessing, we create BigNewsBLN to ensure that all ideologies have almost equal presence to minimize potential bias. POLITICS is not designed to encode potential bias. In Figure 6, the discrepancy in perplexities among different ideologies is more related to the greater coherency among Republicans than Democrats, rather than POLITICS encoding biased knowledge.

In conclusion, there is no greater than minimal risk/harm introduced by either BigNewsBLN or POLITICS. However, to discourage the misuse, we will always warn users that model predictions are for informational purpose only and users should always resort to the broader context to reduce the risk of absorbing biased information.

\(^8\)https://www.copyright.gov/title17/92chap1.html#107.
References


Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In Proceedings of the


Siqi Wu and Paul Resnick. 2021. Cross-partisan discussions on youtube: Conservatives talk to liberals but liberals don’t talk to conservatives.


Appendix A  BIGNEWS Cleaning Steps

In this section, we provide the details of our data cleaning steps for BIGNEWS. We adopt the following cleaning steps to only keep news articles that relate to US politics.

Removing Non-article Pages. Online news websites also post non-news content. We remove such pages by checking their page titles and URLs based on a list of patterns. Sample patterns are shown in Table A1.

Removing Duplicate Pages. We use character-level edit distance to identify duplicate pages. Specifically, we use the following formula to calculate the difference between page $a$ and page $b$:

$$\text{diff}(a, b) = \frac{\text{dist}(a, b)}{\max(\text{len}(a), \text{len}(b))}$$  \hspace{1cm} (5)

where $\text{dist}(a, b)$ is the Levenshtein distance between $a$ and $b$. If the value is less than 0.1, we consider two pages as duplicates and we only keep the one with earlier publication date. Following this procedure, we remove duplicated pages within each media outlet.

Removing Non-politics Pages. To filter out non-politics pages, we build a classifier using training data from BIGNEWS. Since URL typically indicates a page’s topic, we use keywords in the URL to retrieve politics and non-politics training data. The lists of keywords are shown in Table A2. This results in a training dataset with 400,462 politics pages and 310,377 non-politics pages. We also randomly sample 888 pages from the remaining dataset and manually annotate them to use as the test set.

We train a logistic regression model based on unigram and bigram TF-IDF features. To include pages not covered by the lists of keywords in Table A2, we use the trained classifier to classify remaining pages and add those classified with high probability to the training data. This results in a larger training set with 957,424 politics pages and 987,898 non-politics pages. We train the final classifier on the larger training set, and achieve an 88.18% accuracy on the test data.

Removing Non-US Pages. We filter out pages that are not related to US by searching for non-US keywords in the URL. For each of those pages, we only remove it if its text contains no US-related keywords. Examples of keywords used are shown in Table A3.

Removing Media Leaking Phrases. To prevent the model from learning features specific to individual media outlets, we perform a two-step cleaning. First, we mask phrases that mention the media outlet itself (e.g., New York Times, NYTimes, and nytimes.com). Second, we create a list of patterns for frequently appearing sentences (more than 100 times), for each media outlet. For example, as in “author currently serves as a senior political analyst for [MASK] Channel and contributes to all major political coverage”, both the author name and the sentence itself can leak media outlet information. Since sentences with media leaking information usually appear at the beginning or end of the article, we remove any of the first and last two

---

Table A1: Examples of patterns used to filter out pages that are not news stories.

<table>
<thead>
<tr>
<th>Filter Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>Title</strong></td>
</tr>
</tbody>
</table>

Table A2: Full list of keywords used to retrieve positive and negative training data for the politics classifier.

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
</tr>
</tbody>
</table>

Table A3: Examples of keywords used to filter out non-US articles. For text keywords, we collect the names of all presidents, vice presidents, and secretaries of state of US since 2000.

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL Keywords</td>
</tr>
<tr>
<td>Text Keywords</td>
</tr>
</tbody>
</table>

---

We use 0.95 for politics pages and 0.9 for non-politics pages.
<table>
<thead>
<tr>
<th>Media Outlet</th>
<th>Articles (Before Downsample)</th>
<th>Earliest Date</th>
<th>Latest Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Kos</td>
<td>235,244</td>
<td>2009-01-02</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>HuffPost (HPO)</td>
<td>560,581</td>
<td>2000-11-30</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>CNN</td>
<td>152,579</td>
<td>2000-01-01</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>The Washington Post (WaPo)</td>
<td>461,032</td>
<td>2000-01-01</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>USA Today</td>
<td>174,525</td>
<td>2001-01-01</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>The Hill (Hill)</td>
<td>337,256</td>
<td>2002-10-06</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>The Washington Times (TWT)</td>
<td>336,056</td>
<td>2000-01-01</td>
<td>2021-06-30</td>
</tr>
<tr>
<td>Fox News (FOX)</td>
<td>457,550</td>
<td>2001-01-12</td>
<td>2021-06-25</td>
</tr>
<tr>
<td>Breitbart News (Breitbart)</td>
<td>285,530</td>
<td>2009-01-08</td>
<td>2021-06-30</td>
</tr>
</tbody>
</table>

Table A4: Statistics of BigNews corpus. Media outlets are sorted by ideology from left to right.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of steps</td>
<td>2,500</td>
</tr>
<tr>
<td>batch size</td>
<td>2048</td>
</tr>
<tr>
<td>maximum learning rate</td>
<td>0.0005</td>
</tr>
<tr>
<td>learning rate scheduler</td>
<td>linear decay with warmup</td>
</tr>
<tr>
<td>warmup percentage</td>
<td>6%</td>
</tr>
<tr>
<td>optimizer</td>
<td>AdamW (Loshchilov and Hutter, 2019)</td>
</tr>
<tr>
<td>weight decay</td>
<td>0.01</td>
</tr>
<tr>
<td>AdamW beta weights</td>
<td>0.9, 0.98</td>
</tr>
<tr>
<td>$\delta_{ideo}$</td>
<td>0.5</td>
</tr>
<tr>
<td>$\delta_{story}$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table A5: Hyperparameters for continued pretraining.

Appendix B  News Story Alignment

As shown in Equation 1, we combine text similarity and entity similarity to be the final story similarity score. Only title and the first five sentences are considered in the calculation. We further require aligned articles $a$ and $b$ to satisfy two constraints:

- The difference in publication dates of $a$ and $b$ is at most three days.
- $a$ and $b$ must contain at least one common named entity in the title or in the first three sentences.

We use CoreNLP to extract named entities in articles (Manning et al., 2014). For the second constraint, we further apply Crosswikis to map each entity to a unique concept in Wikipedia (Spitkovsky and Chang, 2012). When calculating entity similarity, we split each entity into single words and remove stop words. After alignment, we use the procedure described in Appendix A to remove duplicate articles in the same story cluster. The hyperparameters are $\alpha = 0.4$ and $\theta = 0.23$.

Evaluating Alignment Algorithm. We search the hyperparameters on the Basil dataset (Fan et al., 2019) and test the algorithm on the Allsides dataset collected in Cao and Wang (2021). The Allsides dataset consists of manually aligned news articles from 251 media outlets. After removing media outlets not in BigNewsBLN, we obtain 2,904 articles on 1,316 stories.

To evaluate the performance of the alignment algorithm, we add the evaluation dataset into BigNewsBLN and treat each evaluation article as the anchor article for the alignment algorithm. We use the remaining evaluation articles in the same story as relevant articles, which becomes the target to be identified. The algorithm achieves 0.612 mean reciprocal rank (MRR) on the Basil dataset and 0.679 MRR on the Allsides dataset.

Appendix C  Continued Pretraining and Fine-tuning

C.1 Continued Pretraining

We initialize all variants of POLITICS with a RoBERTa-base model (Liu et al., 2019), which contains about 125M parameters. Our implemen-
Table A6: List of prompts used for stance detection tasks. \textit{p} is the input text, and \{target\} is the target of interest. Verbalizer maps the label (e.g., against) to the token (e.g., negative). Some datasets use a third label (neutral).

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Verbalizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{p} [SEP] The stance towards {target} is [MASK].</td>
<td>negative or positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] It reveals a [MASK] stance on {target}.</td>
<td>negative or positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] The speaker holds a [MASK] attitude towards {target}.</td>
<td>negative or positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] What is the stance on {target}? [MASK].</td>
<td>Negative or Positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] The previous passage [MASK] {target}.</td>
<td>negative or positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] The stance on {target} is [MASK].</td>
<td>negative or positive</td>
</tr>
<tr>
<td>\textit{p} [SEP] The stance towards {target}: [MASK].</td>
<td>opposes or favors</td>
</tr>
<tr>
<td>\textit{p} [SEP] The author [MASK] {target}.</td>
<td>oppose or favor</td>
</tr>
<tr>
<td>\textit{p} [SEP] [MASK]. {target}</td>
<td>No or Yes</td>
</tr>
<tr>
<td>\textit{p} [SEP] [MASK]. {target}</td>
<td>No or Yes</td>
</tr>
</tbody>
</table>

Table A7: Hyperparameters used to fine-tune PLMs.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of epochs</td>
<td>10</td>
</tr>
<tr>
<td>patience</td>
<td>4</td>
</tr>
<tr>
<td>maximum learning rate</td>
<td>0.00001 or 0.00002</td>
</tr>
<tr>
<td>learning rate scheduler</td>
<td>linear decay with warmup</td>
</tr>
<tr>
<td>warmup percentage</td>
<td>6%</td>
</tr>
<tr>
<td>optimizer</td>
<td>AdamW</td>
</tr>
<tr>
<td>weight decay</td>
<td>0.001</td>
</tr>
<tr>
<td>AdamW beta weights</td>
<td>0.9, 0.999</td>
</tr>
<tr>
<td># FFNN layer</td>
<td>2</td>
</tr>
<tr>
<td>hidden layer dimension in FFNN</td>
<td>768</td>
</tr>
<tr>
<td>dropout in FFNN</td>
<td>0.1</td>
</tr>
<tr>
<td>sliding window size</td>
<td>512</td>
</tr>
<tr>
<td>sliding window overlap</td>
<td>64</td>
</tr>
</tbody>
</table>

Table A8: Hyperparameters used to train SVM. \(|D|\) is the number of documents in the training set.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernel</td>
<td>linear</td>
</tr>
<tr>
<td>regularization strength</td>
<td>0.3, 1, or 3</td>
</tr>
<tr>
<td>features</td>
<td>unigram and bi-gram TF-IDF</td>
</tr>
<tr>
<td>minimum document frequency</td>
<td>5</td>
</tr>
<tr>
<td>maximum document frequency</td>
<td>0.7 * (</td>
</tr>
</tbody>
</table>

Table A5 lists the training hyperparameters.

Training Details. For triplet loss objectives, we only consider triplets in each mini-batch. We skip a batch if it contains no triplet. For the MLM objective, we truncate the article if it has more than 512 tokens. When masking entities and sentiment words, we only consider those with at most five tokens. When both triplet loss and MLM objectives are enabled, we adopt alternating training strategy as in Ganin et al. (2016) to apply these two objectives for parameter updates in an alternating manner.

C.2 Fine-tuning

For both ideology prediction and stance detection tasks, we fine-tune each model for up to 10 epochs. We use early stopping and select the best checkpoint on validation set among 10 epochs. For ideology prediction tasks, we follow standard practice of using \texttt{[CLS]} token and feed-forward neural networks (FFNN) for classification. For stance detection tasks, we use prompts to fine-tune PLMs. We curate 11 prompts as shown in Table A6, and select the best prompt based on the performance of RoBERTa. Fine-tuning hyperparameters are listed in Table A7.

For the SVM classifier, we use the implementation of TF-IDF feature extractor and linear SVM classifier in scikit-learn (Pedregosa et al., 2011). The classifier’s hyperparameters are listed in Table A8.
Table A9: Comparison with state-of-the-art results on the original VAST dataset. Best results are in bold.

### Appendix D  Downstream Evaluation Datasets

This section lists more details of the eight datasets used in our downstream evaluation as well as their processing steps.

#### D.1  Ideology Prediction

- **Congress Speech**\(^{11}\) (**CongS**: Gentzkow et al., 2018): We filter out speeches with less than 80 words and use the speaker’s party affiliation as the ideology of the speech.
- **AllSides**\(^{12}\) (**AllS**): We crawl articles from AllSides and use the media outlet’s annotated ideology as that of the article.
- **Hyperpartisan**\(^{13}\) (**HP**: Kiesel et al., 2019): We convert the benchmark into a 3-way classification task by projecting media-level ideology annotations to articles.
- **YouTube** (Wu and Resnick, 2021) contains cross-partisan discussions between liberals and conservatives on YouTube. In our experiments, we only keep controversial comments: 1) A video must have at least 1,500 comments and 150,000 views; 2) A comment must have at least 20 replies. The original dataset annotates users’ ideology on a 7-point scale. We further convert it into a 3-way classification task for left, center, and right ideologies. For the comment-level prediction task on YT (**cmt.**), we use the provided user-level ideology annotation. For user-level prediction on YT (**user**), we concatenate all comments by a user.
- **Twitter** (**TW**: Preo¸tiuc-Pietro et al., 2017): We crawl recent tweets by each user and remove replies and non-English tweets. We assume users’ ideologies do not change after their self-report since prior work has shown that people’s ideology is less likely to change across the political spectrum (Fiorina and Abrams, 2008). We sort all tweets from a user chronologically and concatenate them.

#### D.2  Stance Detection

- **BASIL**\(^{14}\) (**Fan** et al., 2019): We convert the original dataset such that the new tasks are to predict the stance towards a target at two granularities: article (**art.**) and sentence (**sent.**) levels. The targets in the dataset can be a person (e.g., Donald Trump) or an organization (e.g., Justice Department).
- **VAST**\(^{15}\) (**Allaway and McKeown**, 2020) predicts the stance of a comment towards a target. The targets in the dataset are noun phrases covering a broad range of topics (e.g., immigration, home schoolers). We notice the original dataset contains contradictory samples, where the same comment-target pair is annotated with opposite stances, and therefore remove duplicate and contradictory samples.
- **SemEval**\(^{16}\) (**SEval**: Mohammad et al., 2016a) predicts a tweet’s stance towards a target. The dataset contains six targets: Atheism, Climate Change, Feminist, Hillary Clinton, Abortion, and Donald Trump. Notably, the last target is not seen during training, and only appears in testing.

---

\(^{11}\)https://data.stanford.edu/congress_text.


\(^{15}\)https://github.com/emilyallaway/zero-shot-stance.

Table A11: Macro F1 scores on 11 evaluation tasks (average of 5 runs) with standard deviations. Tasks are sorted by text length, short to long, within each group. “All avg” is the average of all 11 tasks. Best results are in bold and second best are underlined. Our models with triplet-loss objectives that outperform RoBERTa are in blue. POLITICS uses Ideology + Story Obj. and Upsamp. Ext. + Sentiment. Results where POLITICS outperforms all baselines are highlighted in red. POLITICS has the smallest standard deviations on 5 tasks, showing its stable performance.

Appendix E Task Property

This section introduces detailed definitions of four properties, based on which we divide tasks into two categories reported in Figure 3.

- Formality: Speech and news genres are considered as formal, and others are informal.
- Training set size: Datasets with more than 2,000 training samples are categorized as large, and small otherwise.
- Document length: Datasets with average document length larger than 500 are treated as “long”, and others are short.
- Aggregation level: If a dataset is a collection of single articles/posts/tweets, then it is in the category of “Single”. If posts are concatenated and aggregated at user level, then it is marked as “User”. Specifically, only YouTube User and Twitter in Table 2 are in the “User” category.

Appendix F Comparison with Previous State-of-the-art Models

Here we compare POLITICS with previous state-of-the-art models on three selected datasets: Hyperpartisan (§5.5), VAST (§F.1), and SemEval (§F.2). §F.3 discusses why direct comparisons are not applicable on the other 5 datasets.

F.1 VAST

POLITICS outperforms all previous state-of-the-art models in the literature as well as the strong RoBERTa baseline, as can be seen in Table A9. Following Allaway and McKeown (2020) and Jarvis and Allaway (2021), \( F_{avg} \) is defined as the macro-averaged F1 over all three classes (favor, against, and neutral). The results are reported on the original VAST dataset which contains contradictory samples, where the same comment-target pairs are annotated with opposite stances so they are counted in both categories.

F.2 SemEval

Table A10 show the results of state-of-the-art models and POLITICS on SemEval. Following Mohammad et al. (2016a), \( F_{avg} \) is defined as the macro-averaged F1 over favor and against classes. State-of-the-art models train separate classifiers, one for each target, thus yield better results than POLITICS. Similar observation is made by Mohammad et al. (2016a), where one single SVM that is trained on all five targets performs worse than five one-versus-rest SVM classifiers.

F.3 Reasons for Inapplicable Comparisons

We are unable to directly compare with existing models on datasets other than Hyperpartisan, VAST, and SemEval for the following reasons:

- The original dataset either is used for different tasks that are not ideology prediction or stance detection: Congress Speech (Gentzkow et al., 2018), BASIL (Fan et al., 2019), and YouTube (Wu and Resnick, 2021).
- The dataset is newly collected (AllSides) or contains newly collected samples (Twitter; Preořuc-Pietro et al., 2017).

Appendix G Visualize Attention Weights
Figure A1: Example 1 for attention visualization. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

Figure A2: Example 2 for attention visualization. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

Figure A3: Example 3 for attention visualization. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

Figure A4: Example 4 for attention visualization. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.