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

Anonymous ACL submission

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 experiments. Our model POLITICS outperforms strong baselines on 8 out of 11 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.

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.

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).

1 Introduction

Ideology is an ubiquitous factor in political science, journalism, and media studies (Mullins, 1972; Freeden, 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 (Preoţiuc-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., 2012a; 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.

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), Naive 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., 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**,[1] 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.

[1] We will release our data and models upon acceptance.

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 counterpartisans (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., 2016a), 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.

In the news domain, PLMs have been primarily used for news factuality prediction (Jwa et al., 2019; Zellers et al., 2019; Kaliyar et al., 2021) and topic classification (Liu et al., 2020; Büyüköz et al., 2020; Gupta et al., 2020) by fine-tuning on task-specific datasets. Few work has investigated using PLMs to discern political ideology evinced in texts. One exception is Baly et al. (2020), where they also leverage a triplet-loss pretraining objective. However, our work is different in at least three aspects. First, our triplet is designed to capture semantic (dis)similarity among articles *on the same story* but of different ideologies, while Baly et al. (2020) choose to compact representations of randomly selected articles, possibly with dramatically different stories, of the same ideology. Thus, their learned representations tend to fit the ideological language used in the pretraining news data, and are not as generalizable to different domains as POLITICS. Second, we introduce a *new* story objective which can effectively prevents the model from overly relying on shortcuts by media boilerplate. This component further improves the generalizability of our model to diverse media. Finally, they rely on a small dataset (35k articles) while our BIGNEWS has more than 3m articles, which we will release for future work in this direction. To the best of our knowledge, we are the first to systematically study and release PLMs for 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 (from 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 **BIGNEWS**, 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, **BIGNEWSBLN**, 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:

$$\mathrm{sim}(p_i, p_j) = \alpha * \mathrm{sim}_t(p_i, p_j) + (1-\alpha) * \mathrm{sim}_e(p_i, p_j) \quad (1)$$

where p_i and p_j are two articles, sim_t is the cosine similarity between TF-IDF vectors of p_i and p_j, 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 α and θ are searched on the Basil dataset (Fan et al., 2019), which contains manually aligned articles.[4] After

[2]We use https://www.allsides.com and https://adfontesmedia.com to decide ideology and https://www.alexa.com/topsites to decide popularity.

[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.

	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

Table 1: Statistics of BIGNEWSBLN. Media outlets are sorted by ideology from left (L), center (C), to right (R) based on AllSides and Media Bias Chart. HPO: Huffington Post; WaPo: The Washington Post; NYT: The New York Times; TWT: The Washington Times. Additional statistics of raw data size before downsampling are in Table A4.

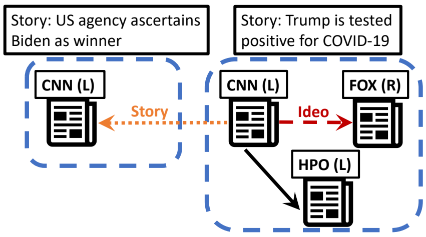

Figure 2: Construction of the ideology and story objectives. The middle CNN article is the anchor in this example. Solid black arrow represents positive-pair relation for both objectives; red dashed arrow denotes negative-pair in ideology objective; orange dashed arrow indicates negative-pair in story objective.

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 (§4.1), which is further augmented by a **story objective** to better focus on content. They are combined with the masked language model objective, which is tailored to focus on entities and sentiments (§4.2), to produce POLITICS (§4.3).

4.1 Ideology-driven Pretraining Objectives

To promote representation learning that better captures ideological content, we leverage BIGNEWSALIGN with articles grouped by stories to provide story-level background for model training. That is, we use triplet loss that operates over **triplets** of <anchor, positive, negative> (Schroff et al., 2015) to encourage anchor and positive samples to have closer representations while contrasting anchor from negative samples.

Our primary pretraining objective, i.e., ideology objective, uses the triplet loss to teach the model to acquire **ideology-informed representations** 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 cluster with the same ideology as *positive* samples, and articles with the opposite ideology as *negative* ones. The ideology objective is formulated as follows:

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

where $\mathcal{T}_{\text{ideo}}$ is the set of all ideology triplets, $\mathbf{t}^{(\mathbf{a})}$, $\mathbf{t}^{(\mathbf{p})}$, and $\mathbf{t}^{(\mathbf{n})}$ are the `[CLS]` representations of anchor, positive, and negative articles in triplet t, δ_{ideo} is a hyperparameter, and $[\cdot]_+$ 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. Similarly, our story objective is formulated as follows:

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

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

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-

Data	Genre	# Train	Len.	Split
Congress Speech (Gentzkow et al., 2018)	speech	7,000	538	rand.
AllSides (*newly collected*)	news	7,878	863	time
BASIL-article (Fan et al., 2019)	news	450	693	story
BASIL-sentence (Fan et al., 2019)	news	1,197	27	story
Hyperpartisan (Kiesel et al., 2019)	news	425	556	rand.
VAST (Allaway and McKeown, 2020a)	cmt	11,545	102	rand.*
YouTube User (Wu and Resnick, 2021)	cmt	1,114	1,213	user
YouTube Cmt (Wu and Resnick, 2021)	cmt	6,832	197	user
SemEval (Mohammad et al., 2016a)	tweet	2,251	17	rand.*
Twitter (Preoţiuc-Pietro et al., 2017)	tweet	1,079	2,298	user

Table 2: Datasets used for evaluating PLMs vary in text genre, training set size (# Train), length, and split criterion. Time split means training on the "past" data and test on the "future". *: splits by the original work.

training work that only considers article-level comparison (Baly et al., 2020).

Concretely, we consider named entities of PERSON, NORP, ORG, GPE and EVENT types. We detect sentiment words using lexicons by Hu and Liu (2004) and Wilson et al. (2005). To focus MLM training more on entities and sentiment, we mask them with a 30% probability, and then randomly mask remaining tokens until 15% (the same probability as used in BERT) of all tokens are reached. We also follow BERT on replacing masked tokens with `[MASK]`, random, and original tokens.

4.3 Overall Pretraining Objective

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

$$\mathcal{L} = \beta * \mathcal{L}_{\text{ideology}} + \gamma * \mathcal{L}_{\text{story}} + (1 - \beta - \gamma) * \mathcal{L}_{\text{MLM}} \quad (4)$$

where $\beta = \gamma = 0.25$. Using $\mathcal{L}$, 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 ~10GB 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 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` [6] (**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 spliting 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;** Preoţiuc-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 (positive, negative, neutral) except for **BASIL (sent.)** that labels positive or negative.

- `BASIL` (Fan et al., 2019) contains news articles with annotations on authors' stances towards entities. **BASIL (sent.)** and **BASIL (art.)** are prediction tasks at sentence and article-levels.
- `VAST` (Allaway and McKeown, 2020a) collects online comments from "Room for Debate", with stances labeled towards the debate topic.
- `SemEval` (Mohammad et al., 2016a) is a shared task on detecting stances in tweets. We consider two setups to predict on seen, i.e. **SEval (seen)**, and unseen, i.e., **SEval (unseen)**, entities.

5.2 Baselines

We consider three baselines. First, we train a linear **SVM** using unigram and bigram features for each

[5] We use `roberta-base` model card from Huggingface.

[6] `https://www.allsides.com`.

	Ideology Prediction							Stance Detection						All avg
	YT (cmt.)	CongS	HP	AllS	YT (user)	TW	Ideo. avg	SEval (seen)	SEval (unseen)	Basil (sent.)	VAST	Basil (art.)	Stan. avg	
SVM	65.34	71.31	61.25	52.51	66.49	42.85	59.96	51.18	32.89	51.08	39.54	30.77	41.09	51.38
BERT	64.64	65.88	48.42	60.88	65.24	44.20	58.21	65.07	40.39	62.81	70.53	45.61	56.88	57.61
RoBERTa	66.72	67.25	60.43	74.75	67.98	48.90	64.34	**70.15**	63.08	68.16	76.25	41.36	63.80	64.09
Our models with triplet loss objective only														
Ideology Obj.	66.20	68.18	64.15	**76.52**	68.15	42.66	64.31	68.78	59.61	64.18	76.03	44.94	62.71	63.58
Story Obj.	66.09	69.11	56.70	74.59	68.89	46.53	63.65	69.02	**63.54**	67.21	76.66	**53.16**	65.92	64.68
Ideology Obj. + Story Obj.	68.91	69.10	63.08	76.23	77.58	**48.98**	67.31	69.66	63.17	64.37	76.18	47.01	64.08	65.84
Our models with masked language model objective only														
Random	67.82	70.32	60.59	73.54	70.77	44.62	64.61	69.16	60.39	69.94	**77.11**	39.16	63.15	63.95
Upsamp. Ent.	**69.06**	70.32	60.09	70.89	71.40	47.16	64.82	69.81	63.08	69.49	76.76	46.46	65.12	64.96
Upsamp. Sentiment	67.41	70.03	56.05	72.35	74.93	48.15	64.82	70.09	60.81	71.28	76.61	44.42	64.64	64.74
Upsamp. Ent. + Sentiment	68.31	**71.42**	58.02	71.90	71.04	47.31	64.67	69.25	62.84	69.23	77.10	43.16	64.32	64.51
POLITICS	67.83*	70.86	**70.25***	74.93	**78.73***	48.92	**68.59**	69.41	61.26	**73.41***	76.73*	51.94*	**66.55**	**67.66**

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 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, *indicating statistical significance (Mann–Whitney U test; Mann and Whitney, 1947, $p \leq 0.05$). Standard deviations (std) are reported in Table A12. The range of std over tasks is $[0.31, 3.42]$ for POLITICS, and $[0.48, 7.35]$ for RoBERTa.

task, since it is a common baseline in political science (Yu et al., 2008; Diermeier et al., 2012b). 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 a sliding window 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 A6) and choose the best one based on the average F1 by RoBERTa on all stance detection tasks:

p `[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 the baselines 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.

Moreover, *our ideology-driven objectives helps 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 on ideology prediction and also improves on stance detection tasks.

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%.

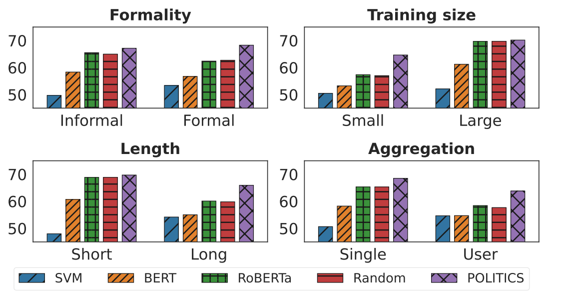

Figure 3: Macro F1 aggregated over tasks of different *formality*, *training size*, *document length* and *aggregation method (single post vs. user posts)*. POLITICS performs better on handling formal language, small training sets, and longer text.

Comparisons with Previous SOTAs. POLITICS achieves an F1 of 77.0 on the *original* `VAST` data where the previous SOTA model obtained 69.2 (Jayaram and Allaway, 2021). Using the *original* binary labels (hyperpartisan or not) on `Hyperpartisan`, POLITICS obtains an accuracy of 85.2, leading SOTA results (Kiesel et al., 2019) by more than 3 points. On `SemEval`, POLITICS obtains an F1 of 71.3 where the best performance is 76.5 by Al-Ghadir et al. (2021). Notably, the comparison includes separate models for different prediction targets, which have been shown to outperform one single classifier (Mohammad et al., 2016a), as done in our setup. Full comparisons with competitive models are included in Appendix F. For other tasks, there is no direct comparison as the datasets are either used for different prediction tasks (e.g., Basil is used for detecting media bias spans) or newly collected.

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 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, and POLITICS consistently outperforms the two counterparts on both tasks, with 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, e.g., when using only 16 samples for fine-tuning on the ideology tasks, than the baselines.

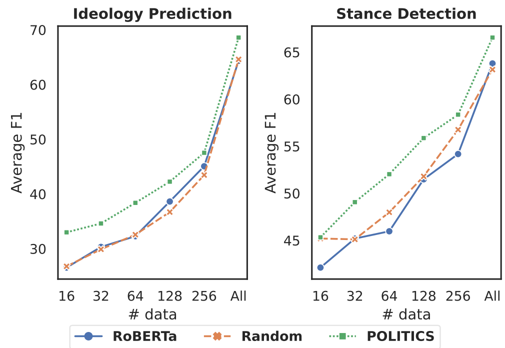

Figure 4: Average of ideology prediction and stance detection performances with few-shot learning. POLITICS uniformly outperforms RoBERTa which is continued pretrained with vanilla MLM (*Random*).

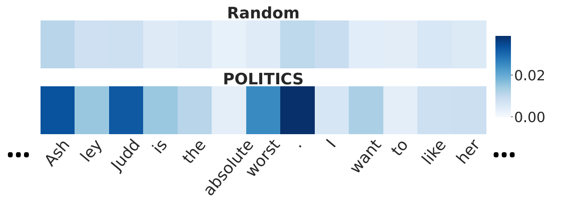

Figure 5: Last layer attention scores between `[CLS]` token and other input tokens (aggregated over all heads). POLITICS captures "worst" and "Ashley Judd". Longer versions of the plot is in Figure A2.

6.2 Ablation Study on POLITICS

We show the impact of removing each ideology-driven pretraining objective and upsampling strategy from POLITICS in Table 4. 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 ideol-

	Ideology Prediction							Stance Detection						All avg
	YT (cmt.)	CongS	HP	AllS	YT (user)	TW	Ideo. avg	SEval (seen)	SEval (unseen)	Basil (sent.)	VAST	Basil (art.)	Stan. avg	
POLITICS	67.83	70.86	70.25	74.93	78.73	**48.92**	68.59	69.41	61.26	**73.41**	76.73	51.94	**66.55**	**67.66**
No Ideology Obj.	-3.78	-2.17	-16.35	-3.28	-12.54	-3.43	-6.93	-0.38	-0.83	-4.22	-0.45	-16.01	-4.38	-5.77
No Story Obj.	**+1.98**	+0.64	-0.72	+0.70	+0.29	-1.78	+0.19	-1.23	**+2.94**	-3.36	-0.87	-10.75	-2.66	-1.11
No Upsamp. Ent.	+0.18	-0.65	-0.05	+0.55	-0.29	-1.20	-0.24	**+0.62**	-0.67	-3.74	-0.55	-1.20	-1.11	-0.64
No Upsamp. Sentiment	+0.75	-0.28	**+0.22**	-1.27	-0.11	-1.40	-0.35	-0.84	+1.67	-3.91	-1.10	**+1.44**	-0.55	-0.44
POLITICS + Ideo. Pred.	+1.46	**+1.10**	-1.01	**+4.72**	**+2.02**	-3.96	**+0.72**	+0.41	-0.52	-3.82	**+0.12**	-3.10	-1.38	-0.23

Table 4: 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.

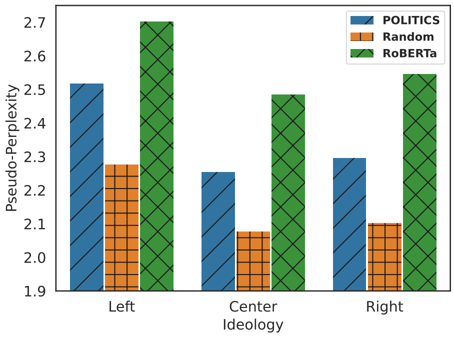

Figure 6: Perplexities of different models on 30K validation articles in BIGNEWSBLN. Perplexities do not drop much on POLITICS, suggesting it can yield superior predictive performance while not overfitting with ideological languages.

ogy prediction since it can directly learn ideology from the annotated labels. However, it has been overfitted to ideology prediction tasks and lacks generalizabilty, 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" and "the worst". 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.

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.

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 and the intellectual property and the privacy rights of the original authors of the texts, i.e., source owners. In the data collection process, the collectors honored privacy rights of article authors and no sensitive information, e.g., identifications, was collected. All participants involved in the process have completed human subjects research training at their affiliated institutions. We also consulted Section 107[7] of the U.S. Copyright Act and ensured that our collection action fell under the fair use category.

8.2 Dataset Usage

BIGNEWS will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.[8] Pretraining corpus details are included in Section 3. The other seven datasets used for downstream evaluation are obtained in the following two ways: **CongS**, **HP**, **BASIL**, **VAST** and **SEval** are obtained by direct download. **CongS** is released under the ODC-BY 1.0 license (free to share, create and adapt); **HP** and **SEval** were released in *ACL shared-task projects, which allows the use of copyrighted material without permission from the copyright holder when it is used for research; For **VAST**, the author explicitly states "We make our dataset and models available for use"; For **BASIL**, we obtain author permission to use the dataset through private correspondence. For **YT** and **TW**, we consult with the corresponding authors and obtain the datasets from them with agreement on not sharing them publicly. We further crawl **AllS** data from the AllSides website while complying with its terms of use. Dataset details are listed in Section 5.1 and Appendix D.

8.3 Benefit and Potential Misuse

Intended use. Assisting the general public to measure ideology of 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 matches how POLITICS would be deployed in real life when handling both ideology prediction and stance detection. We deem that our extensive experiments have covered the major usage of POLITICS.

[7]https://www.copyright.gov/title17/92chap1.html#107
[8]https://creativecommons.org/licenses/by-nc-sa/4.0/

Failure mode is defined as a situation where POLITICS fails to correctly predict the ideology of an individual or an input text. Ideally, the interpretation of our model's prediction should be carried out within the broader context of the input text. However, when taken out of context, prediction results may be misinterpreted by users.
Potential harms. No known harms are observed if POLITICS is being used as intended and functioning correctly. However, if POLITICS malfunctions on stance detection tasks, it could generate opposite results, which might deliver misinformation or make users misunderstand a political figure's stance towards a policy. For vulnerable populations (e.g., people who cannot make the right judgements), the harm might be tremendously magnified if they fail to interpret the ideology prediction and stance detection results in an expected way or blindly trust machine responses.
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 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 one single source.
Potential limitation. Although multiple genres are considered, the genre coverage is not exhaustive, and does not include other trending media for expressing opinions: captions, videos and images. Thus, the predictive performance of POLITICS may still be under investigated. Further, POLITICS is only trained and tested on the same dataset, so its cross-genre ability needs further evaluation.
Bias Mitigation. In our data preprocessing step, we downsample BIGNEWS to BIGNEWSBLN to ensure that each ideology contributes equally to the corpus so as to minimize potential bias. POLITICS is not designed to encode bias. In Figure 6, the discrepancy in perplexities among different ideologies is more related to the greater coherence among Republicans than Democrats (Grossmann and Hopkins, 2016; Benkler et al., 2018), 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.

References

Abdulrahman I. Al-Ghadir, Aqil M. Azmi, and Amir Hussain. 2021. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Inf. Fusion*, 67:29–40.

Emily Allaway and Kathleen McKeown. 2020a. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020b. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Stephen Ansolabehere, Jonathan Rodden, and James M Snyder. 2008. The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102(2):215–232.

Samin Aref and Zachary P Neal. 2021. Identifying hidden coalitions in the us house of representatives by optimally partitioning signed networks based on generalized balance. *Scientific reports*, 11(1):1–9.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 876–885. The Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James R. Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4982–4991. Association for Computational Linguistics.

Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Adam Bonica. 2013. Mapping the ideological marketplace. *ERN: Models of Political Processes: Rent-Seeking*.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, Marseille, France. European Language Resources Association (ELRA).

Shuyang Cao and Lu Wang. 2021. Inference time style control for summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Kakia Chatsiou and Slava Jankin Mikhaylov. 2020. Deep learning for political science. *CoRR*, abs/2005.06540.

Frances Christie and James R Martin. 2005. *Genre and institutions: Social processes in the workplace and school*. A&C Black.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Kareem Darwish, Peter Stefanov, Michaël J. Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 141–152. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012a. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012b. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Benjamin Enke. 2020. What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Robert M. Entman. 2007. Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1):163–173.

Michael Evans, Wayne McIntosh, Jimmy Lin, and Cynthia Cates. 2007. Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4):1007–1039.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6342–6348. Association for Computational Linguistics.

Morris P. Fiorina and Samuel J. Abrams. 2008. Political polarization in the american public. *Annual Review of Political Science*, 11(1):563–588.

Michael Freeden. 2006. Ideology and political theory. *Journal of Political Ideologies*, 11(1):3–22.

Matthew J Gabel and John D Huber. 2000. Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, pages 94–103.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Tim Groseclose, Steven D Levitt, and James M Snyder. 1999. Comparing interest group scores across time and chambers: Adjusted ada scores for the us congress. *American political science review*, 93(1):33–50.

Matt Grossmann and David A Hopkins. 2016. *Asymmetric politics: Ideological Republicans and group interest Democrats*. Oxford University Press.

Shloak Gupta, S Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. 2020. Polibert: Classifying political social media messages with bert. In *Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC.*

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Robert A Hackett. 1984. Decline of a paradigm? bias and objectivity in news media studies. *Critical Studies in Media Communication*, 1(3):229–259.

Kazuaki Hanawa, Shota Sasaki, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2019. The sally smedley hyperpartisan news detector at semeval-2019 task 4. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 1057–1061. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1348–1356. Asian Federation of Natural Language Processing / ACL.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Tim Isbister and Fredrik Johansson. 2019. Dick-preston and morbo at semeval-2019 task 4: Transfer learning for hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 939–943. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan L. Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1113–1122. The Association for Computer Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*

2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 3651–3657. Association for Computational Linguistics.

Sahil Jayaram and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5540–5554. Association for Computational Linguistics.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 840–844. Association for Computational Linguistics.

Heejung Jwa, Dong Bin Oh, Kinam Park, Jang Kang, and Hueiseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multim. Tools Appl.*, 80(8):11765–11788.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Heemin Kim and Richard C Fording. 1998. Voter ideology in western democracies, 1946–1989. *European Journal of Political Research*, 33(1):73–97.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An svm-based approach. *CoRR*, abs/1803.08910.

Michael Laver, Kenneth Benott, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jeffrey Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2021. Voteview: Congressional roll-call votes database.

Jingang Liu, Chunhe Xia, Xiaojian Li, Haihua Yan, and Tengteng Liu. 2020. A bert-based ensemble model for chinese news topic prediction. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering*, New York, NY, USA. Association for Computing Machinery.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14857–14866. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

John Levi Martin. 2015. What is ideology? *Sociologia, Problemas e Práticas*, pages 9–31.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *CoRR*, abs/1605.01655.

Willard A. Mullins. 1972. On the concept of ideology in political science. *American Political Science Review*, 66(2):498–510.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Andrew Peterson and Arthur Spirling. 2018. Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 26(1):120–128.

Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American journal of political science*, pages 357–384.

Anushka Prakash and Harish Tayyar Madabushi. 2020. Incorporating count-based features into pre-trained models for improved stance detection. *CoRR*, abs/2010.09078.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Eitan Sapiro-Gheiler. 2019. Examining political trustworthiness through text-based measures of ideology. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 10029–10030. AAAI Press.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze questions for few shot text classification and natural language inference.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Boris Shor and Nolan McCarty. 2011. The ideological mapping of american legislatures. *American Political Science Review*, 105(3):530–551.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1868–1873. Association for Computational Linguistics.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vertika Srivastava, Ankita Gupta, Divya Prakash, Sudeep Kumar Sahoo, Rohit R. R, and Yeon Hyang Kim. 2019. Vernon-fenwick at semeval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 1078–1082. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2399–2409. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 327–335. ACL.

Andrea Volkens, Tobias Burst, Werner Krause, Pola Lehmann, Theres MatthieÃŸ, Nicolas Merz, Sven Regel, Bernhard WeÃŸels, and Lisa Zehnter. 2021. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2021a.

Marilyn A. Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 592–596. The Association for Computational Linguistics.

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*

Workshop on Methods for Optimizing and Evaluating Neural Language Generation, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016a. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016b. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 384–388. The Association for Computer Linguistics.

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(1):529–544.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097.

Chia-Lun Yeh, Babak Loni, and Anne Schuth. 2019. Tom jumbo-grumbo at semeval-2019 task 4: Hyperpartisan news detection with glove vectors and SVM. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 1067–1071. Association for Computational Linguistics.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.

Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 458–463. The Association for Computer Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

	Filter Patterns
url	/video/, /gallery/, /slideshow/
title	weekly digest, 10 sites you should know, day's end roundup, photos of the week, 5 things you need to know

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

	Keywords
Politics	/politics/, /political/, /policy/, /election/, /elections/, /allpolitics/
Non-politics	/travel/, /sports/, /life/, /movie/, /entertainment/, /science/, /music/, /plated/, /leisure/, /showbiz/, /lifestyle/, /fashion/, /art/,

Table A2: Keywords used to retrieve positive and negative training data for the politics classifier.

Url Keywords	Text US Keywords
/world/, /international/, /europe/, /africa/, /asia/, /latin-america/, /middle-east/	U.S., United States, Obama, Trump, Bush, Biden, Pompeo, Clinton, Pence

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

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 in BigNews.

Removing Non-article Pages. Online news websites also post non-news content. We remove such pages by checking their page title and url, and using a list of patterns to filter out invalid pages. Some example 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:

$$\text{diff}(a, b) = \text{dist}(a, b) / \max(\text{len}(a), \text{len}(b)) \quad (5)$$

where $\text{dist}(a, b)$ is the Levenshtein distance between a and b. If the difference is less than 0.1, we consider two pages as duplicates of each other. For duplicated pages, we only keep the one with the earliest 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 to check whether a page is about politics or not, using training data from BigNews. Since url typically indicates a page's content, 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.

With the training data, we train a unigram and bigram TF-IDF vectorizer to extract features and a logistic regression model for classification. 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 confidence[9] 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 a 88.67% F-1 score and 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-info 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, consider: "`author` currently serves as a senior political analyst for `[MASK]` Channel and contributes to all major political coverage." Both

[9] We use 0.95 for politics pages and 0.1 for non-politics pages.

	# article before downsample	Earliest article	Latest article
Daily Kos	235,244	2009-01-02	2021-06-30
HuffPost (HPO)	560,581	2000-11-30	2021-06-30
CNN	152,579	2000-01-01	2021-06-30
The Washington Post (WaPo)	461,032	2000-01-01	2021-06-30
The New York Times (NYT)	403,191	2000-01-01	2021-06-22
USA Today	174,525	2001-01-01	2021-06-30
Associated Press (AP)	285,685	2000-01-01	2021-06-30
The Hill (Hill)	337,256	2002-10-06	2021-06-30
The Washington Times (TWT)	336,056	2000-01-01	2021-06-30
Fox News (FOX)	457,550	2001-01-12	2021-06-25
Breitbart News (Breitbart)	285,530	2009-01-08	2021-06-30

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

Hyperparameter	Value
number of steps	2,500
batch size	2048
maximum learning rate	0.0005
learning rate scheduler	linear decay with warmup
warmup percentage	6%
optimizer	AdamW (Loshchilov and Hutter, 2019)
weight decay	0.01
AdamW beta weights	0.9, 0.98
δ_{ideo}	0.5
δ_{story}	1.0

Table A5: Hyperparameters for continued pretraining.

the author name and the sentence itself can leak media outlet information. As such sentences usually appear at the beginning or end of the article, we remove any of the first and last two paragraphs, if they contain a sentence that matches any pattern.

Appendix B Story Alignment

As shown in Equation 1, we combine text similarity and entity similarity as the final 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 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 final hyperparameters we use 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 use each evaluation article as the anchor article for the alignment algorithm. We use the remaining evaluation articles in the same story as relevant articles, and the algorithm needs to retrieve them from BIGNEWSBLN. The algorithm achieves 0.612 mean reciprocal rank (MRR) on the Basil dataset and 0.679 MRR on the Allsides dataset.

Prompt	Verbalizer
p `[SEP]` *The stance towards* {target} *is* `[MASK]`.	negative or positive
p `[SEP]` *It reveals a* `[MASK]` *stance on* {target}.	negative or positive
p `[SEP]` *The speaker holds a* `[MASK]` *attitude towards* {target}.	negative or positive
p `[SEP]` *What is the stance on* {target}? `[MASK]`.	Negative or Positive
p `[SEP]` *The previous passage* `[MASK]` {target}.	opposes or favors
p `[SEP]` *The stance on* {target} *is* `[MASK]`.	negative or positive
p `[SEP]` *The stance towards* {target}: `[MASK]`.	negative or positive
p `[SEP]` *The author* `[MASK]` {target}.	opposes or favors
p `[SEP]` `[MASK]` {target}	oppose or favor
p `[SEP]` `[MASK]`. {target}	No or Yes
p `[SEP]` `[MASK]` {target}	No or Yes

Table A6: List of prompts designed for stance detection tasks. p is the input text, and {target} is the target of interest. Verbalizer maps the label (against) to the token (negative) that we want models to predict. Some datasets have a third label (neutral).

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

Table A7: Hyperparameters used to fine-tune PLMs.

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 implementation is based on the HuggingFace transformers library (Wolf et al., 2020).[10] We train each model using 8 Quadro RTX 8000 GPUs for $2,500$ steps. The total training time for POLITICS is 20 hours, and shorter for other variants of it. Table A5 lists the training hyperparameters.

[10]`https://github.com/huggingface/transformers`.

Hyperparameter	Value
kernel	linear
regularization strength	0.3, 1, or 3
features	unigram and bi-gram TF-IDF
minimum document frequency	5
maximum document frequency	$0.7 * \|D\|$

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

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, i.e., training POLITICS, we adopt alternating training as in Ganin et al. (2016) to prime these two objectives to update parameters 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 `[CLS]` token and feedforward 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

Models	VAST		
	F_{favor}	$F_{against}$	F_{avg}
BERT-joint (Allaway and McKeown, 2020b)	54.5	59.1	65.3
TGA Net (Allaway and McKeown, 2020b)	57.3	59.0	66.5
BERT-base (Jayaram and Allaway, 2021)	64.3	58.1	69.2
prior-bin:gold (Jayaram and Allaway, 2021)	64.5	54.6	68.4
RoRBERTa	67.2	71.4	76.5
POLITICS	**68.0**	**72.2**	**77.0**

Table A9: Comparison with state-of-the-art results on original VAST. Best results are in **bold**. Our POLITICS performs best across the board.

Models	Hyperpartisan			
	Acc.	Precision	Recall	F1
Bertha von Suttner (ELMo+CNN; Jiang et al., 2019)	82.2	87.1	75.5	80.9
Vernon Fenwick (Feature Engrg.; Srivastava et al., 2019)	82.0	81.5	82.8	82.1
Sally Smedley (BERT; Hanawa et al., 2019)	80.9	82.3	78.7	80.5
Tom Jumbo Grumbo (SVM; Isbister and Johansson, 2019)	80.6	85.8	73.2	79.0
Dick Preston (ULMFiT Yeh et al., 2019)	80.3	79.3	81.8	80.6
RoRBERTa	84.3	**87.2**	80.6	83.7
POLITICS	**85.2**	86.3	**83.7**	**84.9**

Table A10: Comparison with state-of-the-art results on original Hyperpartisan (with binary labels). Best results are in **bold**. Our POLITICS performs best across the board.

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.

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, right, and center 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;** Preoţiuc-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.

Models	SemEval (Seen)		
	F_{favor}	$F_{against}$	F_{avg}
WKNN (Al-Ghadir et al., 2021)	**84.49**	68.36	**76.45**
PNEM (Siddiqua et al., 2019)	66.56	**77.66**	72.11
MITRE (Zarrella and Marsh, 2016)	59.32	76.33	67.82
pkudblab (Wei et al., 2016b)	61.98	72.67	67.33
SVM-ngrams (Mohammad et al., 2016a)	62.98	74.98	68.98
Majority class (Mohammad et al., 2016a)	52.01	78.44	65.22
BERT	62.89	70.75	66.82
RoRBERTa	67.33	75.52	71.43
POLITICS	67.36	75.29	71.33

Table A11: Comparison with state-of-the-art results on SemEval. Prior work (top panel) trains five models (one per target). On the contrary, in this work, we target a more generalizable approach, i.e., one unified classifier. Due to different problem setups, POLITICS and baselines like RoRBERTa perform worse.

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, 2020a) pre-

[11] `https://data.stanford.edu/congress_text`.

[12] `https://www.allsides.com`.

[13] `https://webis.de/data/pan-semeval-hyperpartisan-news-detection-19.html`.

[14] `https://github.com/marshallwhiteorg/emnlp19-media-bias`.

[15] `https://github.com/emilyallaway/`

	Ideology Prediction							Stance Detection						All avg
	YT (cmt.)	CongS	HP	AllS	YT (user)	TW	Ideo. avg	SEval (seen)	SEval (unseen)	Basil (sent.)	VAST	Basil (art.)	Stan. avg	
SVM	65.34$_{\pm 0.00}$	71.31$_{\pm 0.00}$	61.25$_{\pm 0.00}$	52.51$_{\pm 0.00}$	66.49$_{\pm 0.00}$	42.85$_{\pm 0.00}$	59.96	51.18$_{\pm 0.00}$	32.89$_{\pm 0.00}$	51.08$_{\pm 0.00}$	39.54$_{\pm 0.00}$	30.77$_{\pm 0.00}$	41.09	51.38
BERT	64.64$_{\pm 1.92}$	65.88$_{\pm 1.13}$	48.42$_{\pm 1.44}$	60.88$_{\pm 0.83}$	65.24$_{\pm 1.53}$	44.20$_{\pm 2.03}$	58.21	65.07$_{\pm 1.02}$	40.39$_{\pm 0.53}$	62.81$_{\pm 3.95}$	70.53$_{\pm 0.43}$	45.61$_{\pm 3.92}$	56.88	57.61
RoBERTa	66.72$_{\pm 0.85}$	67.25$_{\pm 0.48}$	60.43$_{\pm 3.13}$	74.75$_{\pm 1.26}$	67.98$_{\pm 4.03}$	48.90$_{\pm 1.53}$	64.34	**70.15**$_{\pm 0.87}$	63.08$_{\pm 0.77}$	68.16$_{\pm 2.55}$	76.25$_{\pm 0.11}$	41.36$_{\pm 7.35}$	63.80	64.09
Our models with triplet loss objective only														
Ideology Obj.	66.20$_{\pm 1.46}$	68.18$_{\pm 0.54}$	64.15$_{\pm 6.82}$	**76.52**$_{\pm 1.62}$	68.15$_{\pm 6.89}$	42.66$_{\pm 10.84}$	64.31	68.78$_{\pm 0.79}$	59.61$_{\pm 3.97}$	64.18$_{\pm 4.41}$	76.03$_{\pm 0.32}$	44.94$_{\pm 5.61}$	62.71	63.58
Story Obj.	66.09$_{\pm 1.05}$	69.11$_{\pm 1.21}$	56.70$_{\pm 2.64}$	74.59$_{\pm 1.68}$	68.89$_{\pm 3.18}$	46.53$_{\pm 3.29}$	63.65	69.02$_{\pm 0.38}$	**63.54**$_{\pm 1.19}$	67.21$_{\pm 2.51}$	76.66$_{\pm 1.29}$	**53.16**$_{\pm 6.76}$	65.92	64.68
Ideology Obj. + Story Obj.	68.91$_{\pm 0.44}$	69.10$_{\pm 0.71}$	63.08$_{\pm 3.10}$	76.23$_{\pm 2.96}$	77.58$_{\pm 2.83}$	**48.98**$_{\pm 1.42}$	67.31	69.66$_{\pm 0.45}$	63.17$_{\pm 1.92}$	64.37$_{\pm 1.58}$	76.18$_{\pm 1.13}$	47.01$_{\pm 7.55}$	64.08	65.84
Our models with masked language model objective only														
Random	67.82$_{\pm 1.30}$	70.32$_{\pm 0.94}$	60.59$_{\pm 2.22}$	73.54$_{\pm 1.55}$	70.77$_{\pm 1.43}$	44.62$_{\pm 2.32}$	64.61	69.16$_{\pm 0.84}$	60.39$_{\pm 0.85}$	69.94$_{\pm 1.61}$	**77.11**$_{\pm 0.53}$	39.16$_{\pm 3.71}$	63.15	63.95
Upsamp. Ent.	**69.06**$_{\pm 1.00}$	70.32$_{\pm 0.39}$	60.09$_{\pm 0.98}$	70.89$_{\pm 1.81}$	71.40$_{\pm 2.23}$	47.16$_{\pm 1.07}$	64.82	69.81$_{\pm 0.61}$	63.08$_{\pm 1.90}$	69.49$_{\pm 1.85}$	76.76$_{\pm 1.01}$	46.46$_{\pm 5.56}$	65.12	64.96
Upsamp. Sentiment	67.41$_{\pm 1.12}$	70.03$_{\pm 0.96}$	56.05$_{\pm 5.68}$	72.35$_{\pm 1.09}$	74.93$_{\pm 2.70}$	48.15$_{\pm 1.30}$	64.82	70.09$_{\pm 0.51}$	60.81$_{\pm 1.22}$	71.28$_{\pm 2.31}$	76.61$_{\pm 0.62}$	44.42$_{\pm 4.91}$	64.64	64.74
Upsamp. Ent. + Sentiment	68.31$_{\pm 0.37}$	**71.42**$_{\pm 0.51}$	58.02$_{\pm 3.34}$	71.90$_{\pm 0.61}$	71.04$_{\pm 3.56}$	47.31$_{\pm 2.07}$	64.67	69.25$_{\pm 0.71}$	62.84$_{\pm 3.93}$	69.23$_{\pm 1.08}$	77.10$_{\pm 0.73}$	43.16$_{\pm 4.95}$	64.32	64.51
POLITICS	67.83$^{*}_{\pm 0.49}$	70.86$_{\pm 0.31}$	**70.25**$^{*}_{\pm 2.10}$	74.93$_{\pm 0.83}$	**78.73**$^{*}_{\pm 1.15}$	48.92$_{\pm 2.19}$	**68.59**	69.41$_{\pm 0.36}$	61.26$_{\pm 1.23}$	**73.41**$^{*}_{\pm 0.97}$	76.73$^{*}_{\pm 0.60}$	51.94$^{*}_{\pm 3.42}$	**66.55**	**67.66**

Table A12: 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 vanilla MLM (Random) are in green. POLITICS uses Ideology + Story Obj. and Upsamp. Ent. + Sentiment. Results where POLITICS outperforms all baselines are highlighted in red, and * indicates statistical significance (Mann–Whitney U test; Mann and Whitney, 1947, $p \leq 0.05$). POLITICS achieves the least standard deviation on five tasks, demonstrating its relative stability.

dicts 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.

Appendix E Task Property

This section introduces detailed definitions of four properties, i.e., how we divide tasks into two categories for each property.

- Formality: Speech and news genres are considered formal, and others are informal.
- Training set size: Datasets with more than 2,000 training samples are considered large, and small otherwise.
- Document length: Datasets with average document length larger than 500 are considered "long", and others are short.
- Aggregation level: If a dataset is a collection of single articles/posts/tweets, then it is categorized as "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 SOTA models

In this section, we discuss compare the performance between POLITICS and existing SOTA models on two selected datasets: `VAST` (§F.1), `Hyperpartisan` (**HP;** §F.2) and `SemEval` (**SEval;** §F.3). §F.4 discusses why direct comparisons are not applicable on the other five datasets.

F.1 VAST

As shown in Table A9, POLITICS outperforms all SOTA models in the literature as well as the strong RoBERTa baseline. Following Allaway and McKeown (2020b); Jayaram 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 pair is annotated with opposite stances.

F.2 Hyperpartisan

Similarly, POLITICS outperforms other models and RoBERTa (except for precision) (see Table A10). Following Kiesel et al. (2019), F1 is defined as the F1 of positive class (i.e., hyperpartisan). The results are reported on the original `Hyperpartisan` dataset with binary labels, i.e., hyperpartisan vs. non-hyperpartisan.

---
zero-shot-stance.

[16] https://alt.qcri.org/semeval2016/task6/index.php?id=data-and-tools.

F.3 SemEval

Table A11 lists the results of existing SOTA models and POLITICS. Following Mohammad et al. (2016a), F_{avg} is defined as the macro-averaged F1 over favor and against classes. Although POLITICS does not outperform existing models, it is mostly because they train five models, one for each target (e.g., Hillary Clinton). Similar pattern is observed in Mohammad et al. (2016a), where one unified SVM (trained on all five targets) performs worse than five one-versus-rest SVM classifiers. However, their "multiple target-dedicate classifier" approach is limited in scalability (to, say, thousands of targets) and generalizability.

F.4 Reasons for Inapplicable Comparisons

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

- The original dataset either is used for different prediction tasks or focuses on different NLP problem: `Congress Speech` (Gentzkow et al., 2018), `BASIL` (Fan et al., 2019), `YouTube` (Wu and Resnick, 2021).
- The dataset is newly collected or contains newly collected samples: `AllSides`, `Twitter` (Preoţiuc-Pietro et al., 2017).

Appendix G Visualize Attention Weights

In this section, we visualize attention weights for more examples.

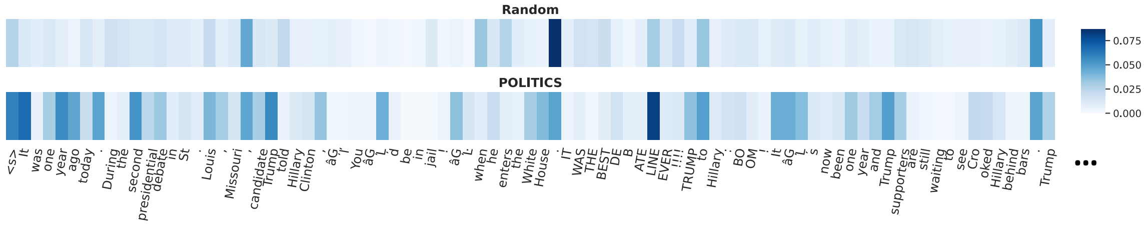

Figure A1: Example 1. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

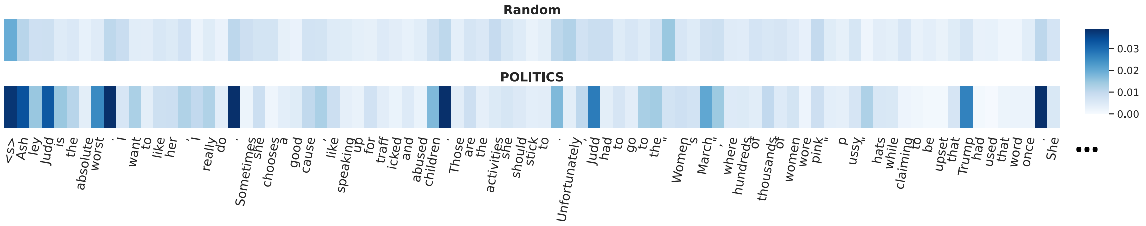

Figure A2: Example 2. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

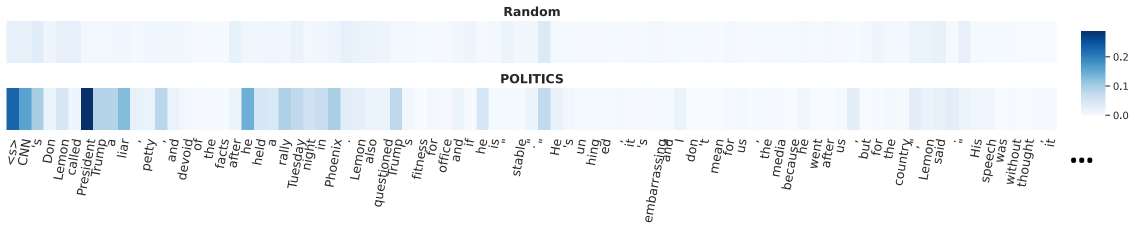

Figure A3: Example 3. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

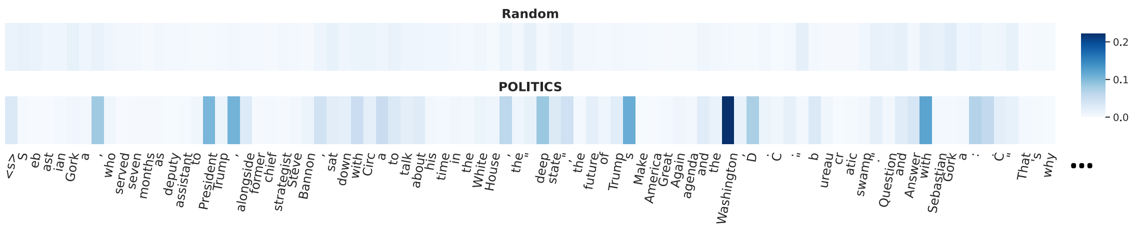

Figure A4: Example 4. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.

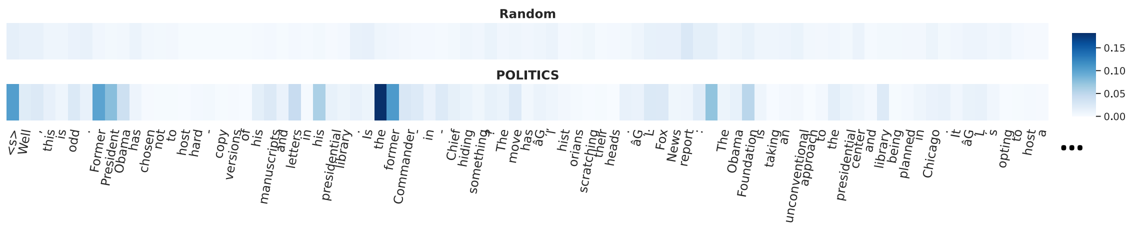

Figure A5: Example 5. Last layer attention weights between [CLS] token and other tokens in the input. We illustrate the first 85 tokens of the article.