A LDA Topic Model

We experimented with a range of hyperparameters to ensure the Latent Dirichlet Allocation (LDA) model was best optimized for our datasets, leveraging the Gensim[4] library. In particular, we removed all stopwords, extremely rare words (tail 10-20% from a unigram distribution), and set the number of topics to 50.

B Self-Ensembling

The core intuition behind consistency regularization is that ensembled predictions are more likely to be correct than single predictions (Laine and Aila, 2017; Tarvainen and Valpola, 2017). To this end, Laine and Aila (2017) introduce a **student** and **teacher** network that yield single predictions and ensembled predictions, respectively.

After learning from labeled samples, the student may produce varying, dissimilar predictions for unlabeled samples due to the stochastic nature of optimization. One potential solution is to ensemble predictions across time to converge at the *most likely* prediction (Laine and Aila, 2017). Tarvainen and Valpola (2017) improve upon this method by showing that ensembling parameters (as opposed to predictions) results in better predictions. Because the teacher's parameters are smoothed with the student's learned parameters at each iteration, the teacher effectively becomes an ensemble of the student across time.

Further, to ensure that the features learned from the labeled samples are compatible with the unlabeled samples, Laine and Aila (2017); Tarvainen and Valpola (2017); French et al. (2018) motivate a consistency-enforcing approach to bring the student and teacher's predictions closer together. In essence, if a feature learned from samples in the labeled domain is incompatible with samples in the unlabeled domain, the consistency (unsupervised) loss penalizes its incompatibility. Therefore, the interplay between these two networks creates a robust, domain-invariant feature space that characterizes both labeled and unlabeled samples (French et al., 2018). A detailed visualization of the training procedure is presented in Figure 1 in the main body of this paper.

[4] https://radimrehurek.com/gensim/

	Political			Non-Political
	AG	PE	IR	
Train	333	8	156	497
Dev	82	1	33	116
Test	125	8	47	208

Table 6: Distribution of train (In-Domain benchmark *only*), dev, test documents in our expert-annotated COHA subcorpus. For political documents, we break down the distribution into American Government (AG), Political Economy (PE), and International Relations (IR).

C NYT Descriptors

We build a list of "political" descriptors in NYT to determine (a) which labels we can or cannot sample non-political documents from; and (b) which descriptors fall under the three areas of political science we consider for our multi-label task (American Government, Political Economy, and International Relations).

Because documents can be tagged with multiple descriptors, we build a list of descriptors whose documents have significant overlap with US POLITICS & GOVERNMENT. The second author, a political science graduate student, filtered this list to 57 descriptors that are political in nature.

For (a), we sample 4,600 non-political documents whose descriptors do not overlap with the 57 political descriptors described above. For (b), the same political science graduate student assigns each descriptor with one or more area labels. We use this label information to build an NYT dataset for our tasks. The 57 political descriptors and their corresponding area labels are tabulated in Table 7.

D Expert-Annotated Dataset

To create an initial COHA subcorpus of 56,000 documents (8,000 per decade), we sample from the following news sources that consistently appear in across decades: Chicago Tribune, Christian Science Monitor, New York Times, Time Magazine, and Wall Street Journal. Note that these NYT articles (up to year 1986) do not appear in the NYT annotated corpus (Sandhaus, 2008) (starting from year 1987), which we used as our source, training dataset.

From this subcorpus, we perform additional steps to create an expert-annotated dataset (§5). Label distributions for our dataset are presented in Table 6. Although political economy (PE) is

severely underrepresented, we experimentally find that these documents have salient features and are not as difficult to classify. In addition, we employ class imbalance penalties to prevent our model from ignoring these documents.

The source dataset (NYT) was already annotated; to ensure label agreement with our target dataset (COHA), we sampled documents from the source dataset and had our political science graduate students label them to compare against the original label. There were minimal problems here—because NYT has fine-grained labels for their documents, the politically-labeled articles were clearly political and vice-versa.

The target datatset (COHA) was divided into halves and each political science graduate student annotated a half. Prior to annotation, they agreed upon a set of rules to minimize bias in the annotation process. In addition, both of them worked side-by-side during all annotation periods, so they were able to ask each other's opinion in case there was confusion. We also took measures to ensure label correctness after annotation was completed. Each political science graduate student sampled a batch of their political and non-political annotations and sent it to the other to evaluate. Again, there was not much disagreement here as the rules decided upon in the beginning were sufficient to cover most edge cases. Quantitatively, Cohen's $\kappa = 0.95$ as calculated on a mutually annotated subset (Cohen, 1960).

	Area Label		
Topic	**AG**	**PE**	**IR**
International Relations			✓
Presidents and Presidency (US)	✓		
Presidential Elections (US)	✓		
War and Revolution			✓
Presidential Election of 2000	✓		
Presidential Election of 2004	✓		
Law and Legislation	✓		
Civil War and Guerrilla Warfare			✓
International Trade and World Market		✓	
Presidential Election of 1996	✓		
Public Opinion			
Economic Conditions and Trends		✓	
Bombs and Explosives			✓
Arms Sales Abroad			✓
United States Economy		✓	
Missiles and Missile Defense Systems			✓
Oil (Petroleum) and Gasoline		✓	
Appointments and Executive Changes	✓		
Foreign Service			✓
Prisoners of War			✓
War Crimes, Genocide and Crimes Against Humanity			✓
Vice Presidents and Vice Presidency (US)	✓		
Arms Control and Limitation and Disarmament			✓
Military Bases and Installations			✓
Presidential Election of 2008	✓		
Whitewater Case	✓		
Vietnam War	✓		✓
Governors (US)	✓		
Energy and Power		✓	
Stocks and Bonds		✓	
State of the Union Message (US)	✓		
Wages and Salaries		✓	
Church-State Relations	✓		
Shiite Muslims			✓
Special Prosecutors (Independent Counsel)	✓		
White House (Washington, DC)	✓		
Federal Taxes (US)		✓	
Illegal Aliens	✓		
Social Security (US)	✓		
Political Prisoners	✓		✓
Watergate Affair	✓		
Government Employees	✓		
Sunni Muslims			✓
Third World and Developing Countries			✓
Customs (Tariff)		✓	
Welfare (US)		✓	
Gun Control	✓		
Global Warming	✓		
Interest Rates		✓	
Vetoes (US)	✓		
Futures and Options Trading		✓	
Attorneys General	✓		
Layoffs and Job Reductions		✓	
Nazi Policies Toward Jews and Minorities			✓
Government Bonds		✓	
Police Brutality and Misconduct	✓		
Executive Privilege, Doctrine of	✓		

Table 7: Political descriptors in NYT. Each descriptor is categorized under one or more political science areas: American Government (AG), Political Economy (PE), and International Relations (IR).