# The Congressional Classification Challenge: Domain Specificity and Partisan Intensity

HAO YAN*, Washington University in St. Louis
SANMAY DAS, Washington University in St. Louis
ALLEN LAVOIE†, Washington University in St. Louis
SIRUI LI, Washington University in St. Louis
BETSY SINCLAIR, Washington University in St. Louis

In this paper, we study the effectiveness and generalizability of techniques for classifying partisanship and ideology from text in the context of US politics. In particular, we are interested in how well measures of partisanship transfer across domains as well as the potential to rely upon measures of partisan intensity as a proxy for political ideology. We construct novel datasets of English texts from (1) the Congressional Record, (2) prominent conservative and liberal media websites, and (3) conservative and liberal wikis, and apply text classification algorithms to evaluate domain specificity via a domain adaptation technique. Surprisingly, we find that the cross-domain learning performance, benchmarking the ability to generalize from one of these datasets to another, is in general poor, even though the algorithms perform very well in within-dataset cross-validation tests. While party affiliation of legislators is not predictable based on models learned from other sources, we do find some ability to predict the leanings of the media and crowdsourced websites based on models learned from the Congressional Record. This predictivity is different across topics, and itself a priori predictable based on within-topic cross-validation results. Temporally, phrases tend to move from politicians to the media, helping to explain this predictivity. Finally, when we compare legislators themselves across different media (the Congressional Record and press releases), we find that while party affiliation is highly predictable, within-party ideology is completely unpredictable. Legislators are communicating different messages through different channels while clearly signaling party identity systematically across all channels. Choice of language is a clearly strategic act, among both legislators and the media, and we must therefore proceed with extreme caution in extrapolating from language to partisanship or ideology across domains.

CCS Concepts: • **Applied computing → Law, social and behavioral sciences**; *Economics*; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Political science; text classification; political ideology; partisanship; domain adaptation

**ACM Reference Format:**
Hao Yan, Sanmay Das, Allen Lavoie, Sirui Li, and Betsy Sinclair. 2019. The Congressional Classification Challenge: Domain Specificity and Partisan Intensity. In *ACM EC '19: ACM Conference on Economics and Computation (EC '19), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3328526.3329582

*Ordering of the authors is alphabetical with sole exception due to significant contributions by Hao Yan.
†Now at Google Brain


*EC '19, June 24–28, 2019, Phoenix, AZ, USA*

ACM ISBN 978-1-4503-6792-9/19/06…$15.00
https://doi.org/10.1145/3328526.3329582

## 1 INTRODUCTION

Political discourse is a fundamental aspect of government across the world, especially so in democratic institutions. In the US alone, billions of dollars are spent annually on political lobbying and advertising, and language is carefully crafted to influence the public or lawmakers [18, 20]. Matthew Gentzkow won the John Bates Clark Medal in economics in 2014 in part for his contributions to understanding the drivers of media "slant". With the increasing prevalence of social media, where activity patterns are correlated with political ideologies [3], companies are also striving to identify users' partisanship based on their comments on political issues, so that they can recommend specific news and advertisements to them.

Typically ideological estimates are generated from legislative roll call data [13, 40] and more recently have been estimated from other sources such as FEC data [8], newspaper accounts [11], and legislative speech records [19]. Since the expansion of ideal point estimation by Poole and Rosenthal [40], ideal point estimates have served as the most ubiquitous explanation for legislative behavior in political science, helping to understand topics that range from the efficacy of particular institutions [2], the rise of polarization [36], and legislative gridlock [33]. The recent rise in measurement strategies for legislative ideology enables new empirical tests of models for legislative behavior, as these fresh data sources provide rich opportunities for comparisons across time and venue.

It is, therefore, unsurprising that measuring partisanship through text has become an important methodological problem in domains including computer science [35, e.g.], political science [27, e.g.], and economics [22, e.g.]. Many methods based on phrase counting, econometrics, and machine learning have been proposed for the problem of classifying political ideology from text [1, 23, 32]. These methods are often used to generate substantive conclusions. For example, Gentzkow et al. [23] report that partisanship in Congress, measured as the ease of telling which party a speaker is from based on a fragment of text they generated, has been increasing in recent years. In any case, it is now clear that it is reasonable in some domains to estimate partisanship from snippets of text. Therefore, it is tempting to extrapolate in two directions. First, to generalize measures of partisanship *across domains*, and second, to measure *partisan intensity*, or *ideology* on a spectrum, using the confidence of the prediction of party membership based on text. An example of the first kind of generalization would be to ask if a machine learning classifier trained on the Congressional Record could successfully classify the partisan leanings of media columnists or individuals on social media. An example of the second would be to ask if a measure of the predicted probability that someone belongs to a specific party (say Democratic) aligns well with a measure of ideology (say the first dimension of the DW-Nominate score that is thought to measure ideology on the "liberal" vs. "conservative" line).

In this paper we extensively test the validity of these two types of extrapolations. For cross-domain generalization, our results are mixed. We compile datasets corresponding to text from legislators (the Congressional Record and press releases), media (opinion and political articles from Salon.com and Townhall.com, as well as additional websites for robustness tests), and crowdsourced collective intelligence (Conservapedia and RationalWiki). We show that it is, in general, very difficult to generalize from one domain to another, even with state-of-the-art domain adaptation techniques from machine learning, with one exception. The exception is that measures based on the Congressional Record have some limited success in classifying articles from the media, consistent with the use of the Congressional Record by Gentzkow and Shapiro [22]. We show that this predictability is driven in part by the temporal movement of phrases from the Congressional Record to the media. Further, there are significant differences in this kind of predictability based on topic. Using LDA [7], we build a topic model on the unlabeled text from both domains, hard-classify articles to their predominant topic, and then build individual classifiers for each topic.

When training on labeled data from the Congressional Record, there are many topics in which the performance is very good, and these topics are identifiable *a priori* by ranking the topics by *within-domain* cross-validation accuracy. Topics with high within-domain and transfer accuracy include tax and health policy, climate change, and foreign policy in the middle east, while topics with low accuracy include abortion, opinions on the media, and text largely filled with procedural phrases. Importantly, the inverse result does not hold – training classifiers with labeled data from the media does not lead to good performance in identifying party affiliation on the Congressional Record.

The second question is whether the probability of party affiliation as measured by text is a good measure of ideology. In US politics, the gold standard for a first approximation to ideology is usually taken to be the first dimension of the DW-Nominate score (for convenience, we refer to this simply as the DW-Nominate score for the rest of this paper), a measure based on voting behavior [40]. It is generally accepted that the DW-nominate score is useful for measuring ideology on a left-right, or liberal-conservative axis. We construct two text-based measures of partisanship, one from the Congressional Record and one from press releases scraped from the websites of members of Congress. We find that, again, predicting party affiliation is easy, with learned classifiers achieving very high accuracy both within and across domains. However, the probability of party affiliation is not a useful measure of where a legislator's ideology falls on a within-party basis. The within-party partisanship scores provide very little information about within-party DW-nominate scores. Further, there is almost no within-party relationship between the text-based scores estimated on the Congressional Record and estimated on press releases.

Taken together, our results are informative about the use of language in political speech and cautionary in terms of how one can use machine learning based measures to identify party affiliation and political ideology. In keeping with recent literature that identifies growing partisan polarization, our text-based measures of party affiliation perform very well at identifying (out-of-sample) which party the author of a piece of text belongs to, as long as the text domain is kept constant. However, the measures are not well correlated with standard measures of ideology within parties, indicating that political speech is likely a different dimension of ideology. The fact that there is little correlation between the text-based measures applied to Congressional floor speeches and to press releases of the same legislator implies that legislators are communicating different messages through different channels while clearly signaling party identity in both.

## 2 RELATED WORK

### 2.1 Political Text Classification and Labeled Data

Political partisanship classification can be a difficult task even for people – only those who have substantial experience in politics can correctly classify the partisanship behind given articles or sentences. In many political labeling tasks, it is even more essential than in tasks that could be thought of as similar (e.g. labeling images, or identifying positive or negative sentiment in text) to ensure that labelers are qualified before using the labels they generate [10, 32]. Gentzkow and Shapiro [22] get around the lack of human-labeled data by directly using text from members of Congress, and labeling this text according to the party affiliation of the speaker.

One of the reasons why classification of political texts for inexperienced people is hard is because different sides of the political spectrum use slightly different terminology for concepts that are semantically the same. For example, in the US debate over privatizing social security, Democrats typically used the phrase "private accounts" whereas Republicans preferred "personal accounts" [22]. Nevertheless, it is recognized that "dictionary based" methods for classifying political text have trouble generalizing across different domains of text [27].

Despite this, it is relatively common in the social science literature to assume that classifiers trained to recognize political partisanship on labeled data from one type of text can be applied to different types of text (e.g. using phrases from the Congressional Record to measure the slant of news media [22], or using citations of different think tanks by politicians to also measure media bias [28]). However, these papers are classifying the bias of entire outlets (for example, *The New York Times* or *The Wall Street Journal*) rather than individual pieces of writing, like articles. Such generalization ability is not obvious in the context of machine learning methods working with smaller portions of text, and must be put to the test.

One question we ask in this paper is whether the increasingly excellent performance of machine learning models in cross-validation settings will generalize to the task of classifying political partisanship in text generated from a *different* source. For example, can a political ideology classifier trained on text from the Congressional Record successfully distinguish between news articles that we would commonly assume to come from Democratic and Republican points of view? This is a particularly interesting question not only from a technical point of view but also from a substantive one. That is, many political texts are written by authors with different incentives, with different concepts of partisanship, and at different points in time (so that one source may use different language and another may mirror that language later, although both reflect similar latent partisanship). By undertaking a set of experiments to evaluate the capacity of these machine learning models to classify across different domains, we effectively evaluate whether partisanship is reflected similarly across domains as well. We assemble three datasets with very different types of political text and an easy way of attributing labels to texts. The first is the Congressional Record, where texts can be labeled by the party of the speaker. The second is a dataset of articles from two popular web-based publications, `Townhall.com`, which features Republican columnists, and `salon.com`, which features Democratic writers. The third is a dataset of political articles taken from Conservapedia (a Republican response to Wikipedia) and RationalWiki (a Democratic response to Conservapedia). In each of these cases there is a natural label associated with each article, and it is relatively uncontroversial that the labels align with common notions of Democrat and Republican.

### 2.2 Partisanship

Political partisanship in U.S. media has been well studied in economics and other social sciences. Groseclose and Milyo [28] calculate and compare the number of times that think tanks and policy groups were cited by mainstream media and Congress members. Gentzkow and Shapiro [22] generate a partisan phrase list based on the Congressional Record and compute an index of partisanship for U.S. newspapers based on the frequency of these partisan phrases. Budak et al. [10] use Amazon Mechanical Turk to manually rate articles from major media outlets. They use machine learning methods (logistic regression and SVMs) to identify whether articles are political news, but then use human workers to identify political ideology in order to determine media bias. Ho et al. [30] examine editorials from major newspapers regarding U.S. Supreme Court cases and apply the statistical model proposed by Clinton et al. [13]. All of the above research gives us quantitative political slant measurements of U.S. mainstream media outlets. However, these political ideology classification results are corpus-level rather than article level or sentence level.

### 2.3 Machine learning and political science

The machine learning community has focused more on the learning techniques themselves. Gerrish and Blei [24] propose several learning models to predict voting patterns. They evaluate their model via cross-validation on legislative data. Iyyer et al. [32] apply recursive neural networks in political ideology classification. They use Convote [44] and the Ideological Books Corpus [29]. They present cross-validation results and do not analyze performance on different types of data. Ahmed and Xing

[1] propose an LDA-based topic model to estimate political ideology. They treat the generation of words as an interaction between topic and ideology. They describe an experiment where they train their model based on four blogs and test on two new blogs. However, political blogs are considerably less diverse than our datasets; since the articles in our datasets are generated in completely different ways (speeches, crowdsourcing and editorials). The results in this paper constitute a more general test of cross-domain political ideology learning.

### 2.4 Domain adaptation for text classification

Cross-domain text classification methods are an active area of research. Glorot et al. [25] propose an algorithm based on stacked denoising autoencoders (SDA) to learn domain-invariant feature representations. Chen et al. [12] come up with a marginalized closed-form solution, marginalized stacked denoising autoencoders (mSDA). Recently, Ganin et al. [21] have proposed a promising "Y" structure end-to-end domain adversarial learning network, which can be applied in multiple cross-domain learning tasks.

### 2.5 Bias, opinion, and partisanship on social media

Cohen and Ruths [14] investigate the classification of political leaning across three different groups (based on activity level) of Twitter users. Without any domain adaptation methodology, they show that cross-domain classification accuracy declines significantly compared with in-domain accuracy. Our work provides a view across much more diverse data sources than just social media, and engages the question of domain adaptation more substantively.

There has also been recent work on identifying information patterns and opinions in collective intelligence venues like Wikipedia [15, 17]. Wikipedia itself is the largest encyclopedia project in the world and is widely used in both natural language processing and political science studies [9, 37]. While Wikipedia aims for a neutral point of view and aims is considered to have become nonpartisan as many users have contributed to political entries [26], there is also evidence that some users try to manipulate content systematically [16], and automated partisanship identification without needing in-domain labeled data would be extremely helpful for this task.

## 3 PREDICTING PARTY IDENTIFICATION ACROSS DOMAINS

### 3.1 Data

Mainstream newspapers and websites have been widely used to estimate political ideology [5, 10, 22]. However, when viewed as data with a latent partisanship, these texts fall short: mainstream newspapers and websites contain many non-political articles, and the political articles in these texts are typically non-partisan [10]. In this project we want texts that are written as explicit political texts: we want texts that are written to communicate about partisanship. Therefore, we identify three sources of data from different data generating processes that we expect to be partisan: (1) The Congressional Record, containing statements by members of the Republican and Democratic parties in the US Congress; (2) News media opinion articles from Salon (a left-leaning website) & Townhall (a right-leaning one); and (3) Articles related to American politics from two collectively constructed "new media" crowd-sourced websites, Conservapedia (Republican) & RationalWiki (Democratic). We rely on these datasets as we anticipate their texts will likely reflect their latent political partisanship. We are particularly interested in the extent to which latent partisanship differs across sources – that is, the extent to which latent partisanship is reflected in the speech of the Congressional Record – a source controlled by the members of Congress – in contrast to the latent partisanship that is reflected in the news articles – a source controlled by the journalists – and in contrast to the crowdsourced "new media" of Conservapedia and RationalWiki – a source

| Year | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Democrat (CR) | 14504 | 11134 | 17990 | 11053 | 14580 | 11080 | 11161 | 8540 | 9673 | 7956 | 0 | 0 |
| Republican (CR) | 11478 | 9289 | 12897 | 8362 | 13351 | 7878 | 9141 | 6841 | 8212 | 6585 | 0 | 0 |
| Salon | 1613 | 1561 | 2161 | 2598 | 2615 | 1650 | 1860 | 1630 | 865 | 123 | 0 | 0 |
| Townhall | 27 | 143 | 290 | 341 | 174 | 176 | 258 | 380 | 441 | 674 | 0 | 0 |
| RationalWiki | 0 | 0 | 302 | 514 | 666 | 854 | 1086 | 1208 | 1342 | 1402 | 1480 | 1480 |
| Conservapedia | 0 | 93 | 1752 | 2381 | 2933 | 3214 | 3467 | 3698 | 3792 | 3863 | 3937 | 3938 |

Table 1. Article distributions by year in the three datasets. Democrat (CR), Salon, and RationalWiki are Democratic, while Republican (CR), Townhall, and Conservapedia are Republican.

controlled arguably by highly-informed political citizens. Each of these data sources comes from a different author with different incentives: we want to establish the extent to which the latent partisanship in each is similar.

Below we briefly describe each of these sources of data.

**Congressional Record.** The U.S. Congressional Record preserves the activities of the House and Senate, including every debate, bill, and announcement. We use party affiliations of speakers as labels. We retrieve the floor proceedings of both the Senate and House from 2005 to 2014. We separate the proceedings into segments with a single speaker. For each of these segments, we extract the speaker and their party affiliation (Democrat, Republican or independent). In order to focus on partisan language, we exclude speech from independents, and from clerks and presiding officers.

**Salon and Townhall.** We collect articles tagged with "politics" from Salon, a website with a progressive/liberal ideology, and all articles from Townhall, which mainly publishes reports about U.S. political events and political commentary from a conservative viewpoint. We test the robustness of our specific choices by also collecting data from several different "partisan" news websites and repeating some of our tests (see the appendix for more details).

**Conservapedia and RationalWiki.** Conservapedia (http://www.conservapedia.com/) is a wiki encyclopedia project website. Conservapedia strives for a conservative point of view, created as a reaction to what was seen as a liberal point of view from Wikipedia. RationalWiki (http://rationalwiki.org/) is also a wiki encyclopedia project website, which was, in turn, created as a liberal response to Conservapedia. RationalWiki and Conservapedia are based on the MediaWiki system. Once a page is set up, other users can revise it. For RationalWiki, we download pages (including redirect pages, which we later remove) ranking in the top 10000 in number of revisions. We further select pages whose categories contain the following word stems: *liber, conserv, govern, tea party, politic, left-wing, right-wing, president, u.s. cabinet, united states senat, united states house.* Because the Conservapedia community has more articles than RationalWiki, we download the top 40000 pages (again, including redirect pages which are later removed). We apply the same political keywords list we use for RationalWiki. We always use the last revision of any page for a given time period.

Table 1 shows the counts of articles in the Democratic and Republican parts of each of the three datasets by year. Our datasets have the following properties that make them useful for partisan evaluation in the context of U.S. politics: (1) The content is selected to be relevant to U.S. politics; (2) The content can predictably be labeled as Democratic or Republican by a somewhat knowledgeable human; (3) The creation times of items in the three datasets have substantial overlap.

### 3.2 Methodology

Our key goal is to establish predictivity of party from text. Modern machine learning techniques are excellent at out-of-sample prediction, and using machine learning to determine how predictable something is from a given set of variables is now becoming accepted practice even in the social sciences [6, 38]. We apply a standard set of machine learning tools for text classification that we now describe.

*3.2.1 Text Preprocessing.* We perform some preprocessing on all the datasets to extract content rather than references and metadata, and also standardize the text by lowercasing, stemming, removing stopwords and other extremely common and venue-specific words.

*3.2.2 Logistic Regression Models.* Logistic regression is a standard and useful technique for text classification. We extract bigrams from the text and Term Frequency-Inverse Document Frequency weighting to construct the feature representation for logistic regression to use (and denote the overall method TF-IDFLR in what follows). We use the implementation provided in the scikit-learn machine learning package [39] with the "balanced" option to deal with the problem of class imbalance.

*Marginalized Stacked Denoising Autoencoders for domain adaptation.* Marginalized Stacked Denoising Autoencoders (mSDA) [12] are a state-of-the-art cross-domain text classification method [21]. Autoencoders can be used as building blocks of deep learning architectures; they can learn useful intermediate representations by being trained to minimize a reconstruction error on (unlabeled) training data. The idea of using *stacked denoising* autoencoders (SDAs) for domain adaptation in text classification is due to Glorot et al. [25]. The denoising idea is to randomly corrupt the input data $x$ as $\tilde{x}$ to prevent memorization. The denoising autoencoder (DA) is then trained to minimize the reconstruction error, $V(x, g(h(\tilde{x})))$, where $V$ is a loss function, and $g(\cdot)$ and $h(\cdot)$ are called the decoder and encoder respectively, with different non-linear activation functions. These denoising autoencoders can then be stacked, with the outputs of encoding layers feeding into the next layer, and the final output of the SDA constitutes a new feature representation of $x$. The hope is that the unsupervised training process can take place on data from multiple domains, even if labels are only available from one of these domains, and the actual classification function can be learned from the new feature representation to the label space.

The marginalized SDA (mSDA) offers tremendous speedups in training with virtually no cost in terms of performance by marginalizing out the noise such that parameters of the model can be obtained in closed-form [12].

We use TF-IDF bag-of-bigrams vectors as the input to mSDA, implemented using the original mSDA Python package[1], in combination with the logistic regressions described above in our domain adaptation experiments.

*3.2.3 Semi-Supervised Recursive Autoencoders.* Recently, there have been rapid advances in text sentiment and ideology classification based on recursive neural networks. Most of this work is based on sentence or phrase level classification. Some of these methods use fully labeled [43] or partially labeled [32] parsed sentence trees, and some need large numbers of parameters [41, 43]. Since we have large datasets available to use, we use semi-supervised recursive autoencoders (RAE) [42], which do not need parse trees, labels for all nodes in the parse trees, or a large number of parameters.

We use the MATLAB package distributed by Socher et al. [42][2]. We do not transform the words down to their linguistic roots when we apply the RAE method since we need to use a word dictionary. RAEs are used only in the domain adaptation experiments.

[1] http://www.cse.wustl.edu/~kilian/code/files/mSDA.zip
[2] http://nlp.stanford.edu/~socherr/codeDataMoviesEMNLP.zip

| Training \ Test | Congressional Record | Salon & Townhall | Conservapedia & RationalWiki |
|---|---|---|---|
| Congressional Record | 0.83 (TF-IDFLR)<br>0.81 (RAE) | **0.69**(mSDA)<br>0.67 (TF-IDFLR)<br>0.59(RAE) | 0.47(mSDA)<br>**0.49** (TF-IDFLR)<br>0.47 (RAE) |
| Salon & Townhall | **0.60**(mSDA)<br>0.59 (TF-IDFLR)<br>0.54 (RAE) | 0.92(TF-IDFLR)<br>0.90(RAE) | 0.52(mSDA)<br>0.51 (TF-IDFLR)<br>**0.55**(RAE) |
| Conservapedia & RationalWiki | **0.53**(mSDA)<br>0.50 (TF-IDFLR)<br>0.47 (RAE) | **0.58**(mSDA)<br>0.53 (TF-IDFLR)<br>0.57 (RAE) | 0.85 (TF-IDFLR)<br>0.82 (RAE) |

Table 2. Domain adaptation test based on three data sets

## 3.3 Experiments

*3.3.1 The failure of cross-domain party identification.* To test the feasibility of learning a model of party identification on one domain and then using it on another, we evaluate our methods on individual articles. We use logistic regression with TF-IDF features and recursive autoencoders as linear / nonlinear classifiers, respectively. Text classification across different domains is a difficult problem due to the different generative distributions of text [12, 21]. We use mSDA as a domain adaptation technique to mitigate the impact of this problem. We also find that variations in language use over time can significantly impact results (see Appendix), so we restrict our methods to train and test only on data from the same year (using five-fold cross validation), and then aggregate results across years.[3]

Table 2 shows the average AUC for each group of experiments. It is interesting to note that the within-domain cross-validation results (on the diagonal) are excellent for both the linear classifier and the RAE.[4] However, the naive cross-domain generalization results are uniformly terrible, often barely above chance. While we could hope that using a sophisticated domain-adaptation technique like mSDA would help, the results do not support this hypothesis: in only one cross-domain task (generalizing from the Congressional Record to Salon and Townhall) does it help to achieve a reasonable level of accuracy.

It is also interesting to note that cross-domain predictivity within *news media*, but across different *media properties* falls somewhere in between, with AUCs around 0.7, a little higher than those that can be achieved in going from the Congressional Record to the news media (see Appendix for details of those results).

*3.3.2 Failure of domain adaptation, or distinct concepts?* There are two plausible hypotheses that could explain these negative results. H1: The domain adaptation algorithm is failing (probably

[3]Some implementation details: For the vectorizer of TF-IDFLR method, we set min_df = 5 and ngram_range = (2, 2). Other parameters are the defaults in scikit-learn package. The parameters setting here are the default for TF-IDF method for all following experiments in this section. The RAE algorithm trains embeddings using sentences subsampled from the data in order to balance conservative and liberal sentences, and then a logistic regression classifier is used on top of the embeddings thus trained. The marginalized stacked denoising autoencoder, which is expected to find features that convey domain-invariant political ideology information, is run on TF-IDF bigram features before a logistic regression is applied on top of that feature representation.

[4]We take care to manually remove all possible extraneous text that could identify the source (e.g. Salon or Townhall), and check the most predictive features in the classifier to ensure that there is no leakage of the class from the text. Results in the Appendix on consistency across time also show that there is a consistent decline in within-domain classification accuracy as the test data is temporally further removed.

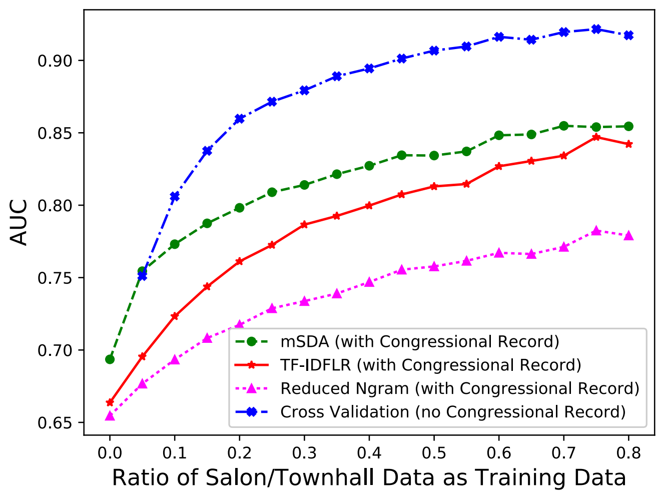


Fig. 1. AUC on Salon/Townhall as a function of the proportion of the labeled (Salon/Townhall) dataset used in training. The results show that including labeled data from the Congressional Record never helps and actively hurts classification accuracy in almost all settings, and that restricting features to ngrams with sufficient support in both datasets does not help either.

because it is easy to overfit labeled data from any of the specific domains), or H2: The specific concepts we are trying to learn are actually different or inconsistent across the different datasets. We perform several experiments to try and provide evidence to distinguish between these hypotheses. First, we may be able to reduce overfitting by restricting the features to ngrams that have sufficient support (operationally, at least 5 appearances) in both sets of data (this reduces the dimensionality of the space and would lead to a greater likelihood of the "true" liberal/conservative concept being found if there were many accurate hypotheses that could work in any individual dataset). Second, we can examine performance as we include more and more *labeled* data from the target domain in the training set. In the limit, if the concepts are consistent, we would not expect to see any degradation in (cross-validation) performance on the source domain from including labeled data from the target domain in training.

We focus on the Salon/Townhall and Congressional Record data sets here since they are the most promising for the possibility of domain adaptation. We combine part of the Salon/Townhall data with Congressional Record as training set. Then we use the rest of the Salon/Townhall data set as the test set, increasing the percentage of the Salon/Townhall dataset used in training from 0% to 80%, and compare with cross-validation performance on just the Salon/Townhall dataset. Figure 1 shows that including labeled data from the Congressional Record never helps and, once we have at least 10% of labels, actively hurts classification accuracy on the Salon/Townhall dataset. Restricting to bigrams that appear in both datasets at least 5 times further degrades the performance. This demonstrates quite clearly that the problem is not overfitting a specific dataset when there are many correct concepts available, it is that the concept of being from Salon or Townhall is significantly different than the concept of being from a Democratic or Republican speech. Therefore, the hope of successful domain-agnostic classification of political party identification based on text data is significantly diminished.

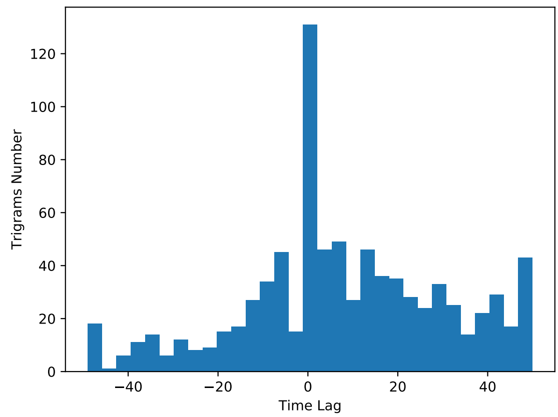


Fig. 2. Distribution of time lag results.

*3.3.3 Generalizing from the Congressional Record.* While our results thus far are mostly negative, we have demonstrated some limited ability to generalize from the Congressional Record to the media dataset. This is in keeping with the corpus level results of Gentzkow and Shapiro [22]. Now we investigate this insight in more depth. We begin by examining the question temporally. Leskovec, Backstrom, and Kleinberg [34] investigated the time lag regarding news events between the mainstream media and blogs. We ask a similar question – who discusses "new" political topics in the first place – Congress or the media?

In order to answer this question, we examine mutual trigrams in the Congressional Record and Salon & Townhall datasets. We find all new trigrams in any given year (those which did not appear in the previous year but appeared at least twice in the media data and five times in the Congressional Record in the given year and the next one), and then construct the time lags between first appearance in each of the two datasets, excluding Congressional recess days.

Figure 2 shows the distribution of these time lags. We only show time lags within 50 days. A positive time lag means those trigrams appear in the Congressional Record sooner than the media. A negative time lag means the media reports those words earlier. The distribution mean is 8.075 and median is 6.0, which demonstrates a tendency for phrases to travel from the Congressional Record to the media rather than the other way round. This helps to explain the relative success of domain adaptation from the Congressional Record to the media dataset, and provides evidence that language moves systematically from legislators to the media.

Second, we turn to a topic analysis. While our hope of successful unconditional domain-agnostic classification of political orientation based on text data was diminished by the above analysis results, we can engage the question on relevant subsets of the data. A straight forward idea is that we first extract text from two sources with the same topic, based on which we can learn a classification model and perform cross domain ideology inference. We perform an experiment where we first use Latent Dirichlet Allocation (LDA) to build a topic model with 40 topics on the combined bag-of-bigrams data from the Congressional Record and Salon/Townhall[5]. Following this we "hard classify" each article in either of the two domains to its predominant topic. We then

[5]We use the Gensim package for LDA model implementation. We set num_topics = 40 and passes = 20 in this experiment.

build individual classifiers within each topic and domain pair, and apply the classifier to articles in the other domain only in the same topic.

Figure 3 shows the main quantitative result of interest. For each domain (CR and ST), first we rank the topics by the cross-validation accuracy achieved within that domain. We show the cross-domain accuracy by taking 5 topics at a time from the top to the bottom of this ranking. In both cases, it is clear that within-domain cross-validation accuracy, especially for the top half of topics, is predictive of the accuracy that can be achieved in the domain adaptation task. The raw numbers make it clear that performance is much better when going from the Congressional Record to Salon and Townhall. Overall, it is clear that the partisan leaning of articles in the news media is highly predictable based on the Congressional Record for some topics, but not for others. An examination of the words associated with these topics (Figure 4) conveys some intuition as to why. The top 5 topics are clearly related to political economy, healthcare and health insurance, evolutionary science and medicine, climate change, and foreign policy, especially in the middle east. On the other hand, the bottom 5 topics (with the exception of Topic 8, which is clearly abortion-related) range from procedural phrases to discussions of political philosophy and specific people. These are interesting observations in terms of the substantive results. The most interesting methodological point here is not just the success of domain adaptation when going from CR to ST on the top 5-10 topics, but also that this success is on the top 5-10 topics based on internal CV accuracy, and therefore this technique can be applied directly without any labels from the transfer domain.

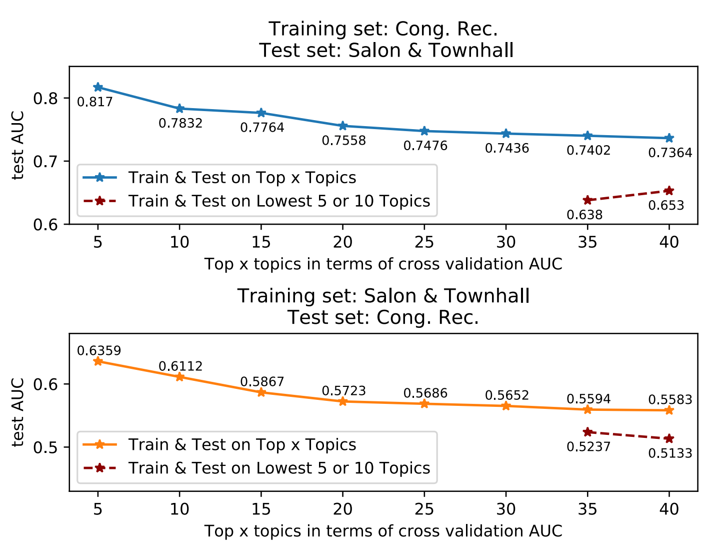


Fig. 3. Congressional Record vs. Salon/Townhall

### 3.4 Discussion

Our results are suggestive along several dimensions. First, it is clear that it would be naive to assume one can generalize party identification from a measure learned on text from one domain to the other. However, it is also clear that there are some patterns of note. In particular, there appears to be a flow of language from legislators to the media, rather than the other way around. Further, predictability on some topics (for example, tax policy) is significantly higher across domains than on others (for example, abortion). The agents who are crafting messages, even in highly partisan

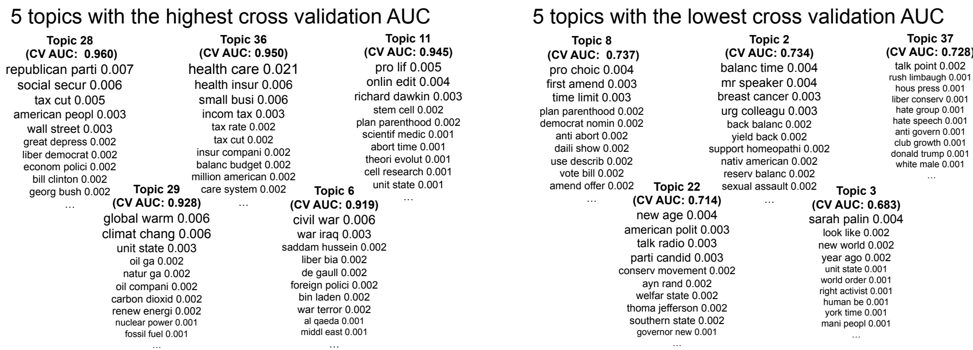


Fig. 4. Bigrams associated with the five most- and least- predictable topics in Congressional Record

domains outside of Congress (e.g. opinion columnists with clear party preferences, or self-identified conservative and liberal Wiki editors) have their own incentives, and the overlap with the language produced by legislators is restricted, albeit likely highly intentional based on the topic analysis.

## 4 PREDICTING PARTISAN INTENSITY

We have thus far developed a text-based measure of party identity, and shown that this at least performs well within domains. Despite the limited success across domains, there is at least evidence in the generalization power of the Congressional Record. We now turn to asking whether it is possible to extrapolate from party prediction to measuring the intensity of a legislator's political ideology, or where they fall on the ideological spectrum.

### 4.1 Data

For the purposes of this section, we focus on members of the House of Representatives from the 113th Congress (2013-2014). In addition to the collection of floor speeches for all of these members, we also collect the 100 latest press releases from their websites (we were only able to gather websites that were available at the time of performing this work). We aggregate the data on a per-legislator basis, because in this case we actively wish to identify how related the measure is when we know it is estimated based on text from the *same person*. In the end, we have 401 representatives in the Congressional Record and 202 representatives in the dataset of press releases (we are not able to crawl all of the representatives' websites due to the diversity in website setups).

### 4.2 Political Ideology Baselines

To evaluate our estimates, we use the following three sources of estimates for political ideology as a baseline. We should flag that these estimates are, in fact, typically considered measurements of political ideology and not necessarily partisanship. Although the two are highly correlated, this will allow us to estimate how partisanship as measured by text and ideology are related.

**DW-Nominate Scores** DW-Nominate scores [40] have been widely used as standard Congressional ideological benchmarks [31]. Each Senator/Representative is scored based on their roll call voting history. The (first dimension) scores range from -1 for extreme liberal to +1 for extreme conservative. We download the DW-Nominate scores from http://voteview.org/.

**DIME Scores** Adam Bonica [8] evaluates Congress members' ideology based on campaign funding sources and proposes the "Database on Ideology, Money in Politics, and Elections" (DIME) baseline. In this model, contributors are assumed to donate based on Congress members' political ideology, thus making it possible to infer a legislator's ideology from the network of donations.

| Training \ Test | Congressional Record | Press Releases |
|---|---|---|
| Congressional Record | 0.9383 | 0.9876 |
| Press Releases | 0.9348 | 0.9783 |

Table 3. Legislator level cross domain test

**Elites Scores** Based on the assumption that Twitter users prefer to follow politicians who share similar political ideology, Pablo Barbera [4] proposes a statistical model that relies upon the network of Twitter users' followee/follower relationships. This allows ideological estimates for all users, both politicians and ordinary users, in a common space, based upon these ties. It is important to note here that legislators do not choose their followers, so that the ideological estimates produced in this matter reflect the followers' preferences.

Each of these sets of estimates are based upon a different data-generating process. For example, members of Congress have control over their roll call votes but not over their Twitter follower network. There is significant variation in these estimates and their comparison with ours is worth serious consideration as we evaluate the extent to which we are observing differences in classification that are associated with domain specificity, with different definitions of partisanship for different actors, or instead with differences in the ways in which partisanship is expressed as a choice in a strategic communication decision.

### 4.3 Experiments

*4.3.1 Partisanship Prediction.* We first test the baseline hypothesis on a per-legislator basis. Is it possible to predict the party membership of a legislator based on text from the Congressional Record or the legislator's press releases. For each of the datasets, we first run "leave-one-out" cross validation[6]. That is, for each representative, we train on data from all other representatives and then test on his/her text. We compare with their true party affiliation and report the AUC. Then we run a cross-domain test, where we train a classification model based on one dataset (Congressional Record or press releases) and test on all legislators in the other. All test AUC results are listed in Table. 3. We see again that party prediction is easy, and we are able to obtain very high accuracy both within- and across- domains.

*4.3.2 Testing a text-based measure of ideology.* Now we turn to the extrapolation task. We follow a similar methodology as above. For each representative, we use all other representatives' speeches in the Congressional Record or press releases (along with party labels) as training data, train a support vector regression (SVR) model with their DW-nominate scores as regression target[7], and then estimate the DW-nominate score of this representative. This score can be considered the legislator's ideology score. Figures 5 and 6 show our main results. These figures plot the relationship between the regression score and the true DW Nominate score, which we take as the gold standard measure of political ideology. Both figures demonstrate that, while the parties are well separated by scores (as above), there is surprisingly little within-party correlation between the text-based measures

[6]For the vectorizer of TF-IDFLR method, the default setting in this section are: min_df = 2, max_df = 500 and ngram_range = (2, 2). Other parameters are the defaults in scikit-learn package.

[7]Again, we use the SVR implementation in scikit-learn package, we use the RBF kernel and set the error term penalty parameter as $C = 10^3$

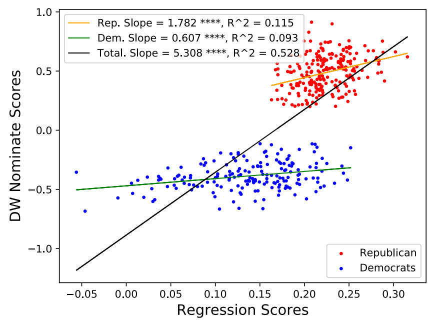


Fig. 5. Text scores based on the Congressional Record versus DW Nominate scores for members of the House of Representatives in the 113th Congress.

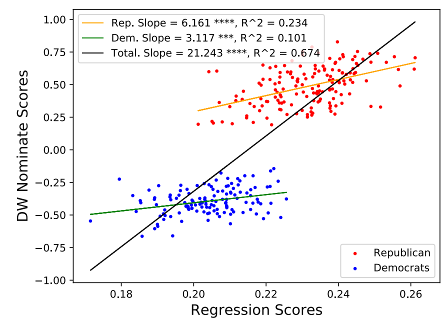


Fig. 6. Text scores based on press releases versus DW Nominate scores for members of the House of Representatives in the 113th Congress.

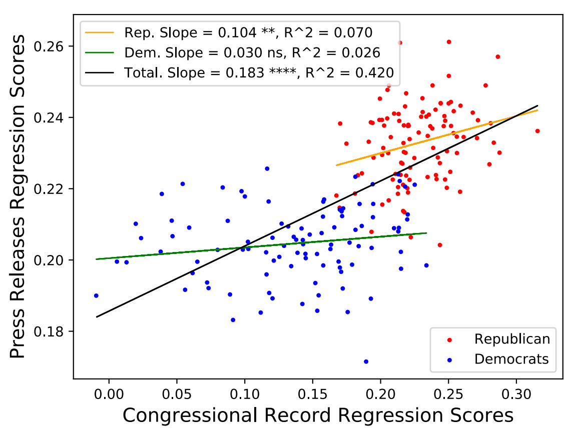


Fig. 7. Text scores based on the Congressional Record versus those based on press releases for members of the House of Representatives in the 113th Congress.

and DW-nominate. Thus, the text-based measures are clearly measuring something quite different from DW-nominate, even though both are good at separating legislators from the two parties.

Perhaps more surprisingly, there is remarkably little correlation (again within-party) among the scores for the same legislator estimated using the Congressional Record versus estimated using press releases (see Figure 7). This is very surprising, considering that the texts are controlled by the same agent (the legislator and their staff). We consider this clear evidence that legislators communicate differently through these channels, and are likely using them to reach different constituencies.

Finally, it is worth comparing our text-based measures with the two recent methods discussed above (DIME scores and elite scores). Figure 8 shows a pattern for DIME scores that is similar to our text-based measures: good party identification but inconsistent within-party estimation of

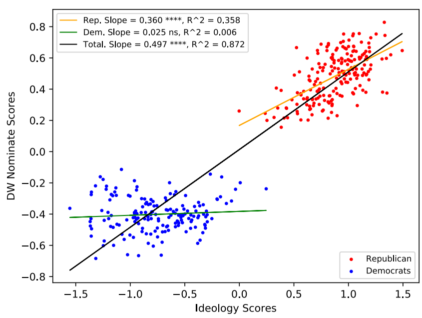


Fig. 8. DIME scores vs. DW Nominate scores for members of the House of Representatives in the 113th Congress.

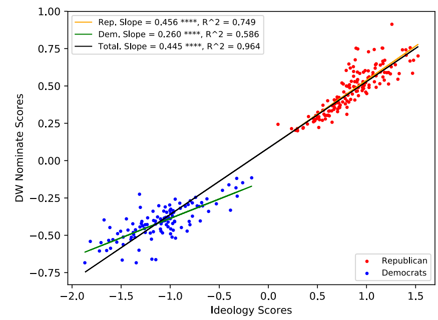


Fig. 9. Estimated Twitter elites ideal scores vs. DW Nominate scores for members of the House of Representatives in the 113th Congress.

ideology.[8] However, Figure 9 shows that the elite scores based on the Twitter network are very consistent with DW-Nominate even within party.

### 4.4 Discussion

Measures of political ideology are tied to theories of how behavior manifests ideology. Spatial models based on roll call votes define the first latent dimension of correlation in voting patterns as ideological behavior. Models based on political networks assume that individuals form ties with people who have similar ideological ideal points. What is important about our measure of ideology is that it adds a new type of behavior, a quantification of the way a legislator presents him- or herself to voters, to the discipline's measures of ideology. The reason multiple measures of ideology are important is that the processes by which ideology is manifested affect what ideology is manifested.

For example, donors and Twitter followers do not vote on legislation. And expressing an opinion on Twitter is not the same as voting on it legislatively. Ideology expressed in one venue is not the same as the ideology expressed through another. Multiple measures of ideology based on different domains of social action makes a broader range of human behavior amenable to analysis. Having measures of ideology in multiple domains like donations and voting offer an increasingly sophisticated set of tools with which to understand how ideology translates across these domains. We establish the relationship between each of these measurements of ideology – some component of ideology captured in a latent space – and party classification from our text-based classifiers.

## 5 CONCLUSIONS

Text analytics is becoming a central methodological tool in analyzing political communication in many different contexts. It is obviously valuable to have a good way of measuring partisanship based on text. Given the success of various methods for party identification based on text, it is tempting to assume that there is enough shared language across datasets that one can generalize from one to the other for new tasks, for example, for detecting bias in wiki editors, or the political orientation of op-ed columnists. It is also tempting to assume that continuous measures of party identification based on text should correlate well with measures of political ideology on a left-right spectrum. Our work sounds a cautionary note in this regard by demonstrating the difficulty of

[8]Within party correlations between the text measures and DIME scores are also limited.

classifying political text across different contexts, and the variability of text-based measures across types of outlets even when they are produced by the same legislator.

We provide strong evidence that, in spite of the fact that writers or speech makers in different domains often self-identify or can be relatively easily identified by humans as Republican or Democrat, the concepts are distinct enough across datasets that generalization is difficult. The one limited exception is that measures estimated using text from the Congressional Record show some promise, especially on a topical basis, in predicting the partisanship of media sources. This is likely because of the temporal movement of phrases from legislative speech to the media. These results suggest that the Congressional Record is not only leading the media in terms of partisan language but moreover that media sources are likely to make different partisan choices.

Second, while polarization is indeed high in the sense that it is easy to predict party affiliations of specific legislators from speech (even across domains), prediction of ideology from speech within party is extremely noisy. In fact, there is almost no correlation of the within-party ideology of a legislator based on his or her Congressional speeches and his or her press releases.

Our overall results suggest that we should proceed with extreme caution in using machine learning (or phrase-counting) approaches for classifying political text, especially in situations where we are generalizing from one type of political speech to another, as the incentives of the authors are not necessarily aligned. Partisanship is a useful lens to different authors at different points in time, and partisan language changes as members of Congress structure their debates. We provide compelling evidence that language moves in a predictable way from legislators to the media.

While we focus in this paper on measures based on predicted probabilities in a classification task, we get qualitatively identical results when measuring political ideology on a real-valued spectrum (the DW-Nominate score [40]) as the target of a regression task (this is only feasible for the Congressional Record, since vote-based scores are available for members of Congress).

The relationship between politicians and their publics continue to evolve as new modes of communication are invented. The trace data documenting these changes are becoming increasingly publicly available in machine-readable formats. However, understanding the methods capable of utilizing this data at scale are required before we can use it to inform our understanding of political behavior. Our project takes one step forward, confirming the validity of partisan classifiers and drawing attention to the heterogeneity with which ideology, particularly within-party ideology, is estimated.

## ACKNOWLEDGMENTS

We thank Roman Garnett, Matthew Gentzkow, Andrew Lo, Christopher Lucas, Jesse Shapiro, Brandon Stewart, and Steven Webster for helpful comments.

## REFERENCES

[1] Amr Ahmed and Eric P Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1140–1150.

[2] Michael A Bailey. 2007. Comparable preference estimates across time and institutions for the Court, Congress, and Presidency. *American Journal of Political Science* 51, 3 (2007), 433–448.

[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[4] Pablo Barbera. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23, 1 (2015), 76.

[5] Matthew A Baum and Tim Groeling. 2008. New media and the polarization of American political discourse. *Political Communication* 25, 4 (2008), 345–365.

[6] Marianne Bertrand and Emir Kamenica. 2018. *Coming apart? Cultural distances in the United States over time.* Technical Report. National Bureau of Economic Research.
[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.
[8] Adam Bonica. 2014. Mapping the ideological marketplace. *American Journal of Political Science* 58, 2 (2014), 367–386.
[9] Adam R Brown. 2011. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics* 44, 2 (2011), 339–343.
[10] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.
[11] Barry C Burden, Gregory A Caldeira, and Tim Groseclose. 2000. Measuring the ideologies of US senators: The song remains the same. *Legislative Studies Quarterly* (2000), 237–258.
[12] Minmin Chen, Zhixiang Xu, Kilian Q Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the International Conference on Machine Learning*. 767–774.
[13] Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98, 2 (2004), 355–370.
[14] Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy!. In *Proceedings of the 7th International AAAI Conference on Web and Social Media*.
[15] Sanmay Das and Allen Lavoie. 2014. Automated inference of point of view from user interactions in collective intelligence venues. In *Proceedings of the International Conference on Machine Learning*. 82–90.
[16] Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)* 10, 4 (2016), 24:1–24:25.
[17] Sanmay Das and Malik Magdon-Ismail. 2010. Collective wisdom: Information growth in wikis and blogs. In *Proceedings of the ACM Conference on Electronic Commerce*. 231–240.
[18] Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics* 122, 3 (2007), 1187–1234.
[19] Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and ideology in Congress. *British Journal of Political Science* 42, 1 (2012), 31–55.
[20] Robert M Entman. 1989. How the media affect what people think: An information processing approach. *The Journal of Politics* 51, 2 (1989), 347–370.
[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.
[22] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78, 1 (2010), 35–71.
[23] Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2019. Measuring polarization in high-dimensional data: Method and application to Congressional speech. *Econometrica* (2019). Forthcoming.
[24] Sean Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning*. 489–496.
[25] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*. 513–520.
[26] Shane Greenstein and Feng Zhu. 2012. Is Wikipedia biased? *American Economic Review* 102, 3 (2012), 343–348.
[27] J. Grimmer and B. M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* (2013), 1–31.
[28] Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120, 4 (2005), 1191–1237.
[29] Justin H Gross, Brice Acree, Yanchuan Sim, and Noah A Smith. 2013. Testing the Etch-a-Sketch hypothesis: A computational analysis of Mitt Romney's ideological makeover during the 2012 Primary vs. General Elections. In *APSA Annual Meeting*.
[30] Daniel E Ho, Kevin M Quinn, et al. 2008. Measuring explicit political positions of media. *Quarterly Journal of Political Science* 3, 4 (2008), 353–377.
[31] Kosuke Imai, James Lo, and Jonathan Olmsted. 2016. Fast estimation of ideal points with massive data. *American Political Science Review* 110, 4 (2016), 631–656.
[32] Mohit Iyyer, Peter Enns, Jordan 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*, Vol. 1. 1113–1122.
[33] Keith Krehbiel. 2010. *Pivotal politics: A theory of US lawmaking*. University of Chicago Press.

[34] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 497–506.

[35] Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 17–32.

[36] Nolan McCarty, Keith T Poole, and Howard Rosenthal. 2016. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

[37] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*. http://arxiv.org/abs/1301.3781

[38] Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.

[39] 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 (October 2011), 2825–2830.

[40] Keith T Poole and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford University Press.

[41] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1201–1211.

[42] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 151–161.

[43] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1631–1642.

[44] Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 327–335.

## APPENDIX

### Consistency across time

The words used to describe politics change across time, as do the topics of importance. Therefore, political articles that are distant in time from each other will be less similar than those written during the same period. Here we study whether this is a significant issue for the logistic regression methods by focusing on the Salon and Townhall dataset.

We repeatedly train on 2 years worth of data[9], and test on 5 future years, one by one. We repeat this process moving forward in time, and then aggregate all the 1-year-out, 2-years-out, 3-years-out, etc. results. So for example, for the 1st round, we train on 2005 and 2006 data, and then test each year from 2007 (mark as $y_1$) to 2011 (mark as $y_5$). The last round involves training on 2008 and 2009 data, and then testing on data from 2010 through 2014.

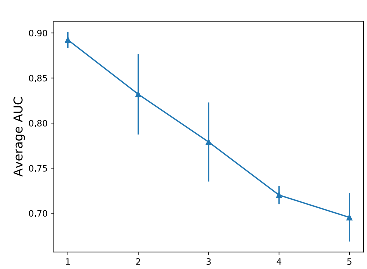


Fig. 10. Salon & Townhall year-based timeline test. The training set is two continuous year Salon & Townhall data. The test sets are next five individual years data.

Figure 10 shows the AUC across time, estimated in this manner. The AUC for $y_1$ is around 0.8923, which means that the Salon & Townhall articles in the immediately following year are similar enough for successful generalization in the ideology classification problem. However, the

[9]For the experiments in this Appendix, we use the TF-IDFLR method. For the vectorizer here, we set min_df = 3, and ngram_range = (2, 2). Other parameters are the defaults.

prediction accuracy goes down significantly as the dates of the test set become further out in the future, as the nature of the discourse changes.

### Ideology generalization across media

We are interested in a robustness test to assess generalizability within the news media, but across different news sources. Therefore we add two liberal media sources (NY Magazine and The Guardian) and two conservative media sources (Breitbart and Fox News) rated clearly as such by allsides.com. We crawl and download (U.S.-related) political news articles from these websites. Table 4 shows the number of articles for each year we scraped from each of these sites. Our approach is to

| Year | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|---|---|---|
| NY Magazine | 1477 | 1619 | 2831 | 1987 | 1011 | 930 | 2326 | 2824 | 2603 |
| Breitbart | 341 | 431 | 6009 | 1384 | 7651 | 7448 | 11240 | 11632 | 7872 |
| Fox News | 1349 | 1338 | 829 | 1316 | 2714 | 3510 | 3722 | 3950 | 2746 |
| The Guardian | 0 | 0 | 0 | 0 | 140 | 952 | 2317 | 2047 | 1153 |

Table 4. Articles scraped from four additional news media websites

construct new pairs of Democrat/Republican training and test sets and measure predictivity in these cross-domain tests. The first experiment is *Salon & Townhall vs. Breitbart & NY Magazine*. We first train a TFIDF-LR model based on the Salon & Townhall data set, and test on Breitbart & NY Magazine[10]. We then switch training and test sets, with data from 2010-2014. The cross-domain train & test results are shown in Table 5. We repeat the above experiment for *Breitbart & NY Magazine vs. The Guardian & Fox News* for data from 2015-2018. Results are shown in Table 6. As we can see, the performance is substantially better than in tests across different sources (with the exception of generalization from the Congressional Record to news media). We also provide 80/20 cross-validation baselines for Breitbart vs. NY Magazine (2010~2014)[11] and The Guardian vs. Fox News (2015~2018) tasks in Table 7. As we can see, the cross-validation AUC is close to that achieved in the Salon vs. Townhall task (0.92) – see Table 2.

| Training Set | Test Set | Avg. AUC |
|---|---|---|
| Breitbart & NY Magazine | Salon & Townhall | 0.706 |
| Salon & Townhall | Breitbart & NY Magazine | 0.700 |

Table 5. Cross media ideology generalization test: *Salon & Townhall vs. Breitbart & NY Magazine*

| Training Set | Test Set | Avg. AUC |
|---|---|---|
| Breitbart & NY Magazine | The Guardian & Fox News | 0.725 |
| The Guardian & Fox News | Breitbart & NY Magazine | 0.702 |

Table 6. Cross media ideology generalization test: *Breitbart & NY Magazine vs. The Guardian & Fox News*

| Data Set | Avg. AUC |
|---|---|
| Breitbart vs. NY Magazine | 0.920 |
| The Guardian vs. Fox News | 0.934 |

Table 7. Media ideology cross-validation test: *Breitbart vs. NY Magazine* and *The Guardian vs. Fox News*

[10] For the vectorizer, we set min_df = 2, max_df = 500, and ngram_range = (2, 2). Other parameters are the defaults.

[11] We also use the full data set from 2010 to 2018. The average AUC is 0.919.