[1]

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 (now at Google Brain)
SIRUI LI, Washington University in St. Louis
BETSY SINCLAIR, Washington University in St. Louis

Political communication often takes complex linguistic forms. Understanding partisanship from text is an important methodological task in studying political interactions between people and in understanding the role of political institutions in both new and traditional media. Therefore, there has been a spate of recent research that either relies on, or proposes new methodology for, the classification of partisanship from text data. Yet, little research has evaluated the robustness of these tools. In this paper, we study the effectiveness of these techniques for classifying partisanship and ideology in the context of US politics. In particular, we are interested in measures of partisanship across domains as well as the potential to rely upon measures of partisan intensity as a proxy for political ideology.

We construct three different datasets of Republican and Democratic English texts from (1) the Congressional Record, (2) prominent conservative and liberal media websites, and (3) conservative and liberal wikis. We apply text classification algorithms to evaluate the possibility of domain specificity via a domain adaptation technique. Our results are surprisingly negative with respect to comparisons with the Congressional Record. We find that the cross-domain learning performance, benchmarking the ability to generalize from one of these datasets to another, is poor when the Congressional Record is compared to media and crowd-sourced estimates, even though the algorithms perform very well in within-dataset cross-validation tests. That is, it is very difficult to generalize from one domain to another. We also find some ability to generalize from the Congressional Record to the media, and show that this ability is different across topics, and 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. These results suggest the need for extreme caution in interpreting the results of machine learning methodologies for classification of political text across domains.

Moreover, these sources have strong correspondence to reliable benchmarks typically used in political science research to estimate political partisanship – that is, predicting party affiliation is easy. However, the within-party scores provide very little information about the within-party ideology of legislators, as gleaned from the canonical benchmarks. We argue this poor performance is evidence for differences in the concepts that generate the true labels across datasets, rather than to a failure of domain adaptation methods. That is, in all of these spaces we are estimating a latent proclivity towards partisan and ideological speech. The incentives of legislators to place partisan and ideological speech in the Congressional Record may be associated with position taking while the incentives of prominent media websites may be to provide novel narratives around partisan and ideological speech. Moreover, roll call voting is a partisan act – an instance where a party has insisted on a recorded vote, typically as a method to enforce party discipline. It is roll call voting that generates the DW-Nominate estimates. That these estimates, in the end, differ from the within-party estimates from the partisan classification in the Congressional Record is likely driven by differences in the incentives of the authors. That these sources do or do not align then, is not necessarily surprising. Legislators are communicating different messages through different channels while clearly signaling party identity systematically across all channels.

[1]Ordering of the authors is alphabetical with sole exception due to significant contributions by Hao Yan.

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 [DellaVigna and Kaplan, 2007, Entman, 1989]. 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 [Bakshy et al., 2015], 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 (Poole and Rosenthal 1997; Clinton, Jackman and Rivers 2004) and more recently have been estimated from other sources such as FEC data (Bonica 2014), newspaper accounts (Burden, Caldeiria and Groseclose 2000),and legislative speech records (Diermeir et al 2012). Since the expansion of ideal point estimation by Poole and Rosenthal (1997), 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 (Bailey 2007), the rise of polarization (McCarty, Poole and Rosenthal 2006), and legislative gridlock (Krehbiel 1998). 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 [Lin et al., 2008, e.g.], political science [Grimmer and Stewart, 2013, e.g.], and economics [Gentzkow and Shapiro, 2010, e.g.]. Many methods based on phrase counting, econometrics, and machine learning have been proposed for the problem of classifying political ideology from text [Ahmed and Xing, 2010, Gentzkow et al., 2017, Iyyer et al., 2014]. These methods are often used to generate substantive conclusions. For example, Gentzkow et al. ([2017]) 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), 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 ([2010]). 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 [Blei et al., 2003], 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 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 [Poole and Rosenthal, 1997]. 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 perhaps 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 [Budak et al., 2016, Iyyer et al., 2014]. Gentzkow and Shapiro ([2010]) 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" [Gentzkow and Shapiro, 2010]. Nevertheless, it is recognized that "dictionary based" methods for classifying political text have trouble generalizing across different domains of text [Grimmer and Stewart, 2013].

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 [Gentzkow and Shapiro, 2010], or using citations of different think tanks by politicians to also measure media bias [Groseclose and Milyo, 2005]). 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 ([2005]) calculate and compare the number of times that think tanks and policy groups were cited by mainstream media and Congress members. Gentzkow and Shapiro ([2010]) 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. ([2016]) 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. ([2008]) examine editorials from major newspapers regarding U.S. Supreme Court cases and apply the statistical model proposed by Clinton et al. ([2004]). 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 ([2011]) propose several learning models to predict voting patterns. They evaluate their model via cross-validation on legislative data. Iyyer et al. ([2014]) apply recursive neural networks in political ideology classification. They use Convote [Thomas et al., 2006] and the Ideological Books Corpus [Gross et al., 2013]. They present cross-validation results and do not analyze performance on different types of data. Ahmed and Xing ([2010]) 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. ([2011]) propose an algorithm based on stacked denoising autoencoders (SDA) to learn domain-invariant feature representations. Chen et al. ([2012]) come up with a marginalized closed-form solution, mSDA. Recently, Ganin et al. ([2016]) 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 ([2013]) 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 [Das and Lavoie, 2014, Das and Magdon-Ismail, 2010]. Wikipedia itself is the largest encyclopedia project in the world and is widely used in both natural language processing and political science studies [Brown, 2011, Mikolov et al., 2013]. 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 [Greenstein and Zhu, 2012], there is also evidence that some users try to manipulate content systematically [Das et al., 2016], and automated partisanship identification without needing in-domain labeled data would be extremely helpful for this task.

3 EXTRAPOLATION PARTY IDENTIFICATION ACROSS DOMAINS

3.1 Data

Mainstream newspapers and websites have been widely used to estimate political ideology [Baum and Groeling, 2008, Budak et al., 2016, Gentzkow and Shapiro, 2010]. 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 [Budak et al., 2016]. 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"

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	6271	5975
Townhall	27	143	290	341	174	176	258	380	441	674	2521	2184
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.

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 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.

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 of Republican by a somewhat knowledgeable human; (3) The creation times of items in the three datasets have substantial overlap.

3.2 Methodology

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 [Pedregosa et al., 2011] with the "balanced" option to deal with the problem of class imbalance.

Marginalized Stacked Denoising Autoencoders for domain adaptation. Marginalized Stacked Denoising Autoencoders (mSDA) [Chen et al., 2012] are a state-of-the-art cross-domain text classification method [Ganin et al., 2016]. Given bag-of-words input of text from two different domains, mSDA provides a closed-form representation of the input, and is faster than the original Stacked Denoising Autoencoder (SDA) [Glorot et al., 2011] without loss of classification accuracy. We use TF-IDF bag-of-bigrams vectors as the input to mSDA, the original mSDA Python package[1] for the implementation of mSDA 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 [Socher et al., 2013] or partially labeled [Iyyer et al., 2014] parsed sentence trees, and some need large numbers of parameters [Socher et al., 2012, 2013]. Since we have large datasets available to use, we use semi-supervised recursive autoencoders (RAE) [Socher et al., 2011], 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. ([2011])[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.

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 [Chen et al., 2012, Ganin et al., 2016]. We use the marginalized stacked denoising autoencoders (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]

[1] http://www.cse.wustl.edu/~kilian/code/files/mSDA.zip

[2] http://nlp.stanford.edu/~socherr/codeDataMoviesEMNLP.zip

[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

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

Table 2. Domain adaptation test based on three data sets

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. 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 are disappointing: 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.

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 algorithm is failing (probably 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

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.

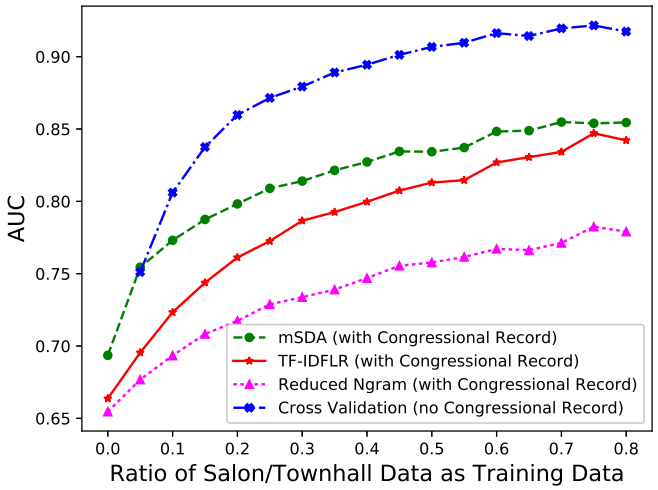

Fig. 1. AUC on Salon/Townhall as a function of the proportion of the labeled (Salon/Townhall) dataset used in training. The results show that including labeled data from the Congressional Record never helps and actively hurts classification accuracy in almost all settings, and that restricting features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

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.

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 [2010]. Now we investigate this insight in more depth. We begin by examining the question temporally. Leskovec, Backstrom, and Kleinberg [2009]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 and 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. Positive time lag means those trigrams appear in the Congressional Record sooner than the media. Negative time lag means media reports those words earlier. The distribution mean is 8.075, median is 6.0 and standard deviation is 22.7682, which show us a definite tendency for phrases to travel from the Congressional Record to the media rather than the other way round, and help to explain the relative success of domain adaptation from the Congressional Record to the media dataset, and provide evidence that language moves systematically from legislators to the media.

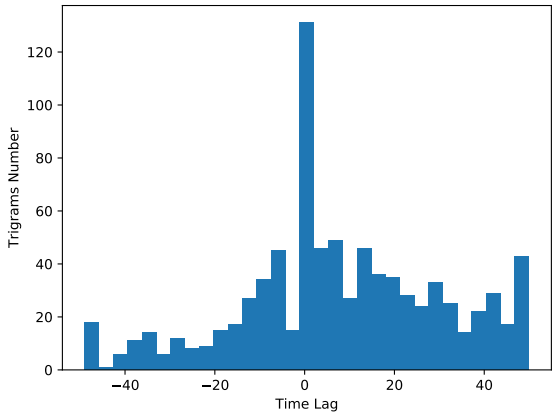

Fig. 2. Distribution of time lag results.

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[4]. Following this we "hard classify" each article in either of the two domains to its predominant topic. We then 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

[4]We use gensim package for LDA model implementation. We set $num\_topics = 40$ and $passes = 20$ in this experiment.

accuracy, and therefore this technique can be applied directly without any labels from the transfer domain.

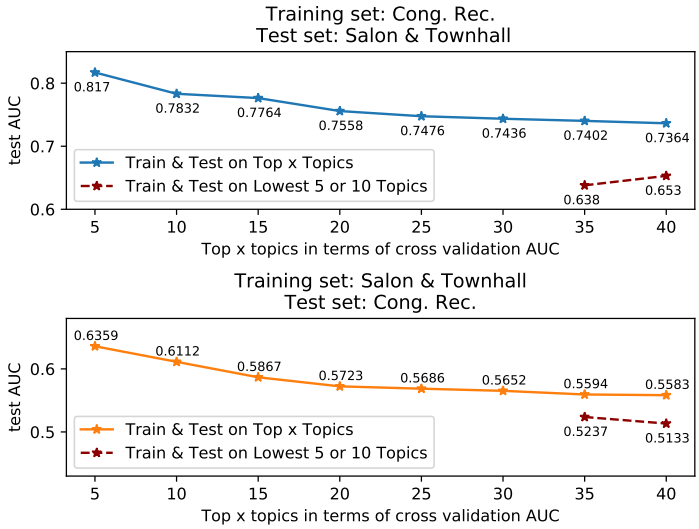

Fig. 3. Congressional Record vs. Salon/Townhall

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 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 EXTRAPOLATING TO 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 are currently available). We aggregate the data on a per-legislator basis, because in this case

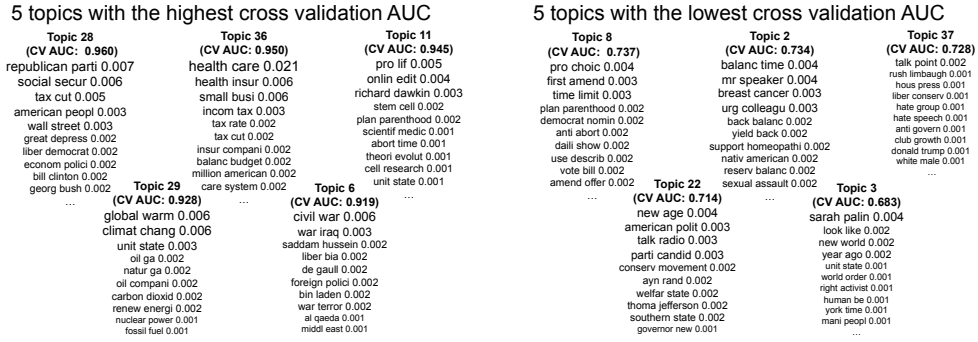

Fig. 4. Bigrams associated with the five most- and least- predictable topics in Congressional Record

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 [Poole and Rosenthal, 1997] have been widely used as standard Congressional ideological benchmarks [Imai et al., 2016]. 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/. We note that this provides DW-Nominate scores through the 114th Congress.

DIME Scores Adam Bonica [2014] 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.

Elites Scores Based on the assumption that Twitter users prefer to follow politicians who share similar political ideology, Pablo Barbera [2015] 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.

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

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[5]. 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 logistic regression model, and then estimate the probability that this representative belongs to the Democratic or Republican party. This probability can be considered the legislator's Classification (CR or PR) score. Figures 5 and 6 show our main results. These figures plot the relationship between the Classification score and the legislator's 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 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 of good party identification but inconsistent within-party identification of ideology for DIME scores, similar to our text-based measures.[6] However, Figure 9 shows that the elite scores based on the Twitter network are very consistent with DW-Nominate even within party.

[5]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.

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

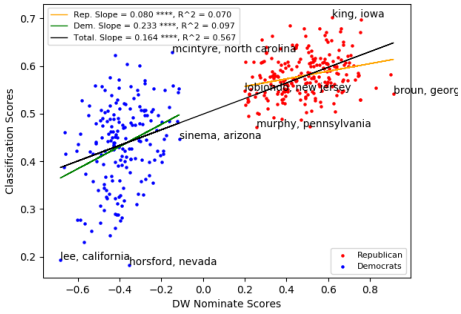

Fig. 5. Text scores based on the Congressional Record versus DW Nominate scores for members of the House of Representatives in the 113th Congress.

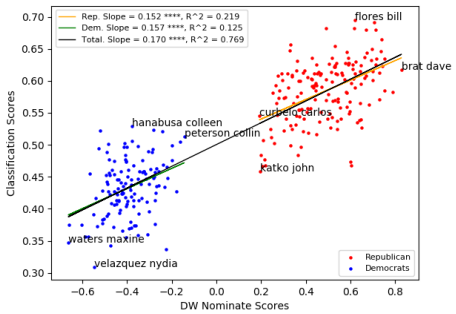

Fig. 6. Text scores based on press releases versus DW Nominate scores for members of the House of Representatives in the 113th Congress.

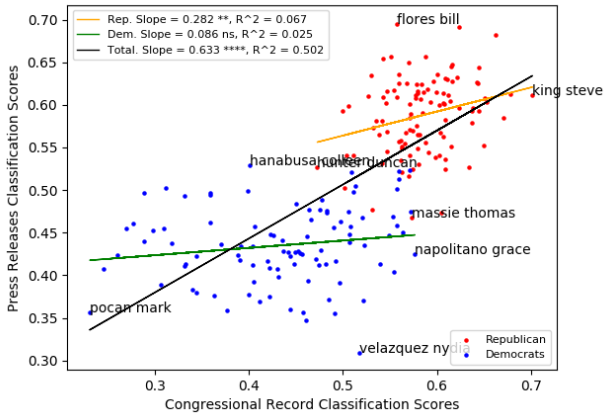

Fig. 7. Text scores based on the Congressional Record versus those based on press releases for members of the House of Representatives in the 113th Congress.

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 a 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

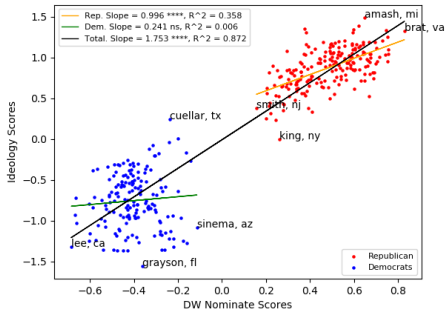

Fig. 8. DIME scores vs. DW Nominate scores for members of the House of Representatives in the 113th Congress.

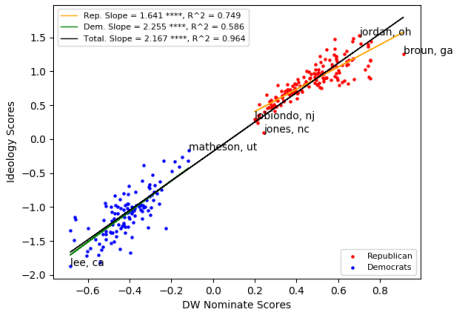

Fig. 9. Estimated Twitter elites ideal scores vs. DW Nominate scores for members of the House of Representatives in the 113th Congress.

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 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.

We note that, 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 [Poole and Rosenthal, 1997]) 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.

REFERENCES

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.

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

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

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

Adam Bonica. 2014. Mapping the ideological marketplace. *American Journal of Political Science* 58, 2 (2014), 367–386.

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.

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.

Minmin Chen, Z Xu, Kilian Q Weinberger, and Fei Sha. 2012. Marginalized stacked denoising autoencoders for domain adaptation. In *Proceedings of the Learning Workshop, Utah, UT, USA*. 767–774.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98, 2 (2004), 355–370.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: Itfis not easy!. In *Proc. ICWSM*.

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.

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.

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.

Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News Effect: Media Bias and Voting. *The Quarterly Journal of Economics* 122, 3 (2007), 1187–1234.

Robert M Entman. 1989. How the media affect what people think: An information processing approach. *The journal of Politics* 51, 2 (1989), 347–370.

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.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2017. *Measuring polarization in high-dimensional data: Method and application to congressional speech.* Technical Report. National Bureau of Economic Research.

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.

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.

Shane Greenstein and Feng Zhu. 2012. Is Wikipedia biased? *American Economic Review* 102, 3 (2012), 343–348.

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.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120, 4 (2005), 1191–1237.

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*.

Daniel E Ho, Kevin M Quinn, and others. 2008. Measuring explicit political positions of media. *Quarterly Journal of Political Science* 3, 4 (2008), 353–377.

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.

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.

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.

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.

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

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.

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

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.

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.

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.

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. We now study whether this is a significant issue for the logistic regression methods by focusing on the Salon and Townhall dataset.

We use 2006 Salon and Townhall articles as a training set and future years (from 2007 to 2014) as separate test sets.[7] Figure 10 shows the AUC across time. The AUC for 2007 is 0.872, which means that the Salon & Townhall articles in 2006 and 2007 are similar enough for successful generalization of the ideology classification problem from one to the other. However, the 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.

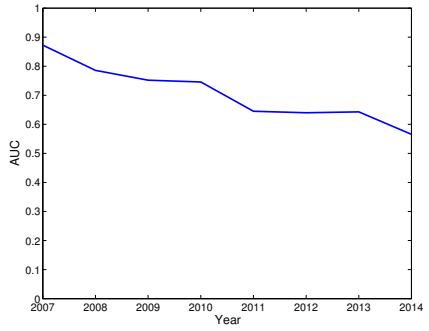

Fig. 10. Salon & Townhall year-based timeline test. The training set is 2006 Salon & Townhall data. The test sets are individual year data from 2007 to 2014, also from Salon & Townhall.

[7]We use the TF-IDFLR method. For the vectorizer, we set $min\_df = 5$ and $ngram\_range = (2, 2)$. Other parameters are the defaults.