The Perils of Classifying Political Orientation From Text

Hao Yan, Allen Lavoie*, and Sanmay Das

Washington University in St. Louis, St. Louis, USA
{haoyan,allenlavoie,sanmay}@wustl.edu

Abstract. Political communication often takes complex linguistic forms. Understanding political ideology from text is an important methodological task in studying political interactions between people 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 political ideology from text data. In this paper, we study the effectiveness of these techniques for classifying ideology in the context of US politics. We construct three different datasets of conservative and liberal 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 with a domain adaptation technique. Our results are surprisingly negative. We find that the cross-domain learning performance, benchmarking the ability to generalize from one of these datasets to another, is poor, even though the algorithms perform very well in within-dataset cross-validation tests. We provide evidence that the poor performance is due to differences in the concepts that generate the true labels across datasets, rather than to a failure of domain adaptation methods. Our results suggest the need for extreme caution in interpreting the results of machine learning methodologies for classification of political text across domains. The one exception to our strongly negative results is that the classification methods show some ability to generalize from the congressional record to media websites. We show that this is likely because of the temporal movement of the use of specific phrases from politicians to the media.

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 [10, 11]. 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 [2], companies are also striving to identify users' ideologies based on their comments on political issues, so that they can recommend specific news and advertisements to them.

The manner in which political speech is crafted and words are used creates difficulties applying standard methods. Political ideology classification is a difficult task

* Now at Google Brain

even for people – only those who have substantial experience in politics can correctly classify the ideology behind given articles or sentences. In many political ideology 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 [5, 21].

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, liberals typically used the phrase "private accounts" whereas conservatives preferred "personal accounts" [13]. Nevertheless, it is well-recognized that "dictionary based" methods for classifying political text have trouble generalizing across different domains of text [17].

Many methods based on machine learning techniques have also been proposed for the problem of classifying political ideology from text [1, 21, 23]. The training and testing process typically follows the standard validation rules: first split the dataset into a training set and a test set, then propose an algorithm and train a classification model based on the training set and finally test on the test set. These methods have been achieving increasingly impressive results, and so it is natural to assume that classifiers trained to recognize political ideology on labeled data from one type of text can be applied to different types of text, as has been common in the social science literature (e.g. Gentzkow and Shapiro using phrases from the Congressional Record to measure the slant of news media [13], or Groseclose and Milyo using citations of different think tanks by politicians to also measure media bias [18]). 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.

The main 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 ideology in text generated from a *different* source. For example, can a political ideology classifier trained on text from the congressional record successfully distinguish between liberal and conservative news articles? One immediate problem we face in engaging this question is the absence of large datasets with political ideology labels attached to individual pieces of writing. Therefore, 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 conservative columnists, and `salon.com`, which features liberal writers. The third is a dataset of political articles taken from Conservapedia (a conservative response to Wikipedia) and RationalWiki (a liberal 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 liberal and conservative. We show that standard classification techniques can achieve high performance in distinguishing liberal and conservative pieces of writing in cross-validation experiments on these datasets.

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 Wikipedia editors, or the political orientation of op-ed columnists. However, is it really reasonable to extrapolate from any of these datasets to others? As a motivating example, we show that the results of training bag-of-bigram linear classifiers using the three different datasets above and then using them to identify the political biases of Wikipedia administrators leads to wildly inconsistent results, with virtually no correlation between the partisanship rankings of the administrators based on the three different training sets. More generally, we show that, with one exception, the unaltered cross-domain performance of different classifiers on these datasets is abysmal, and there is only marginal benefit from applying a state-of-the-art domain adaptation technique (marginalized stacked denoising autoencoders [6]). The exception is in using data from the congressional record to predict whether articles are from Salon or Townhall, consistent with Gentzkow and Shapiro's results on media bias. A temporal analysis suggests that this is because phrases move in a rapid and predictable way from the congressional record to the news media. However, even in this domain, we provide evidence that the underlying concepts (Salon vs. Townhall compared with Democrat vs. Republican) are significantly different: adding additional labeled data from one domain actively hurts performance on the other. Our results are robust to using regressions on measures of political ideology (DW-Nominate scores [26]) rather than simple classifications of partisanship. 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.

1.1 Related Work.

While our methods and results are general, we focus in this paper on political ideology in the US context, since there is already a rich literature on the topic, as well as abundant data. Political ideology in U.S. media has been well studied in economics and other social sciences. Groseclose *et al.,* [18] calculate and compare the number of times that think tanks and policy groups were cited by mainstream media and congresspeople. Gentzkow *et al.,* [13] 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.,* [5] 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.,* [20] examine editorials from major newspapers regarding U.S. Supreme Court cases and apply the statistical model proposed by Clinton *et al.,* [7]. 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.

The machine learning community has focused more on the learning techniques themselves. Gerrish *et al.,* [14] propose several learning models to predict voting patterns. They evaluate their model via cross-validation on legislative data. Iyyer *et al.,* [21]

apply recursive neural networks in political ideology classification. They use Convote [30] and the Ideological Books Corpus [19]. They present cross-validation results and do not analyze performance on different types of data. Ahmed *et al.,* [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.

Cross-domain text classification methods are an active area of research. Glorot *et al.,* [15] propose an algorithm based on stacked denoising autoencoders (SDA) to learn domain-invariant feature representations. Chen *et al.,* [6] come up with a marginalized closed-form solution, mSDA. Recently, Ganin *et al.,* [12] have proposed a promising "Y" structure end-to-end domain adversarial learning network, which can be applied in multiple cross-domain learning tasks.

Cohen *et al.,* [8] 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.

2 Data and Methods

2.1 Data

Mainstream newspapers and websites have been widely used in political ideology research [3, 5, 13]. However, these datasets contain many non-political articles, and the political articles in news datasets are typically non-partisan [5]. Therefore, we carefully construct three datasets 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 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" websites, Conservapedia (conservative) & RationalWiki (liberal). Details of the construction process and the resulting corpora are in the appendix.

2.2 Methods

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.

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 [25].

Marginalized Stacked Denoising Autoencoders for domain adaptation Marginalized Stacked Denoising Autoencoders (mSDA) [6] are a state-of-the-art cross-domain text classification method [12]. 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) [15] 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.

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 [29] or partially labeled [21] parsed sentence trees, and some need large numbers of parameters [27, 29]. Since we have large datasets available to use, we use semi-supervised recursive autoencoders (RAE) [28], 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.,* [28][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.

3 Results

3.1 Cross-domain consistency

The first question is whether training on different domains yields consistent results in classifying political ideology. We evaluate this on a motivating task that is exactly the type of task that one may wish to use these types of tools for, determining ideological bias among Wikipedia administrators.

For each of the 500 most active Wikipedia administrators, we concatenate all the strings they have added to pages on Wikipedia related to U.S. politics and classify the resulting "body of work" of that administrator using the three different training sets (the Congressional Record is #1, Salon/Townhall is #2, and RationalWiki/Conservapedia is #3). Each classifier produces a ranking of these 500 administrators. Shockingly we find that these rankings have **virtually no correlation with each other** (see Table 1).

Somewhat more anecdotally, we can also look at the ranks of some users from each method. We select the three most liberal users according to each of the three classifiers and find their positions in the other two lists. The results are in Table 2 and again demonstrate how diverse the rankings can be based on the training sets.

3.2 Consistency across time

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

User Sorted Lists	Spearman's ρ	Kendall's τ
U_1, U_2	-0.004588	-0.003469
U_2, U_3	0.005201	0.002133
U_3, U_1	-0.073204	-0.048652

Table 1: Correlation between the user ideology ranks as determined by the three different training sets. U_1 is the rank vector based on the classifier trained on Congressional Record, U_2 is based on Salon / Townhall and U_3 is based on RationalWiki / Conservapedia. Both ρ and τ are close to 0, demonstrating almost no correlation (the statistics range from -1 for perfectly anti-correlated to +1 for perfectly correlated).

User Name	U_1	U_2	U_3
Barek	1	282	487
ERcheck	2	387	35
Widr	3	345	496
James086	262	1	356
Penwhale	455	2	300
Dave souza	97	3	240
Gyrofrog	425	141	1
Smartse	416	358	2
Rigadoun	418	38	3

Table 2: Rankings of the three most liberal users according to classifiers trained on each of the training sets.

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.

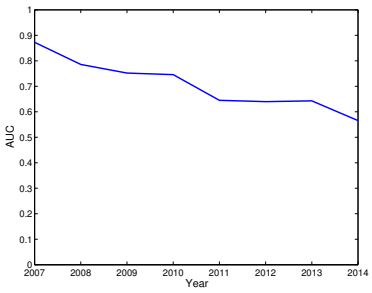

Fig. 1: Salon & Townhall year-based timeline test. The training set is 2006 Salon & Townhall data. The test sets are individual year data from 2007 to 2014, also from Salon & Townhall.

Figure 1 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 classifier 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. It is now clear that our classification methods have generalization problems both across domains and across time.

3.3 Domain adaptation

Now we turn to a more comprehensive analysis. We examine the performance of several different methods across the three labeled datasets. We study linear classifiers and recursive autoencoders as described above, as well as the mSDA method for domain adaptation. In order to account for the effects of time-varying language use demonstrated above, we restrict our methods to train and test only on data from the same year, and then aggregate results across years.

Training Set \ Test Set	Congressional Record	Salon & Townhall	Conservapedia & RationalWiki
Congressional Record	0.8299 (TF-IDFLR) 0.8136 (RAE)	**0.6935**(mSDA) 0.6731 (TF-IDFLR) 0.5937(RAE)	0.4729(mSDA) **0.4940** (TF-IDFLR) 0.4655 (RAE)
Salon & Townhall	**0.6038**(mSDA) 0.5861 (TF-IDFLR) 0.5363 (RAE)	0.9193(TF-IDFLR) 0.9041(RAE)	0.5234(mSDA) 0.5080 (TF-IDFLR) **0.5527**(RAE)
Conservapedia & RationalWiki	**0.5260**(mSDA) 0.5012 (TF-IDFLR) 0.4674 (RAE)	**0.5835**(mSDA) 0.5282 (TF-IDFLR) 0.5711 (RAE)	0.8493 (TF-IDFLR) 0.8180 (RAE)

Table 3: Domain adaptation test based on three data sets

Table 3 shows the average AUC for each group of experiments. 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. The AUC score gaps between cross-validation and domain adaptation results indicate that, even with a state-of-the-art domain adaptation algorithm, cross-text domain political ideology identification is not, at this point, able to give reliable results. It is of note that the best performance is in generalizing from the congressional record to a media dataset (Salon/Townhall) because it adds weight to the existing line of research starting from Gentzkow and Shapiro on how language flows from politicians to the media. (Implementation details and parameter choices for Sections 3.1-3.3 can be found in the appendix)

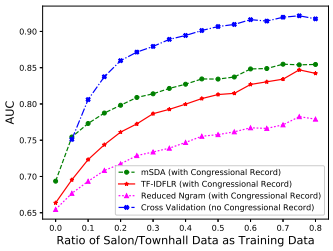

Fig. 2: AUC on Salon/Townhall as a function of the proportion of the labeled (Salon/Townhall) dataset used in training. The results show that including labeled data from the Congressional Record never helps and actively hurts classification accuracy in almost all settings, and that restricting features to ngrams with sufficient support in both datasets does not help either.

3.4 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 2 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 orientation based on text data is significantly diminished.

3.5 Temporal movement of topics

The silver lining so far is that there is at least some ability to predict the political orientation of web-based news media based on the congressional record. We can further investigate this insight and demonstrate the utility of the data we have collected by examining the question temporally. Leskovec *et al.,* [22] 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?

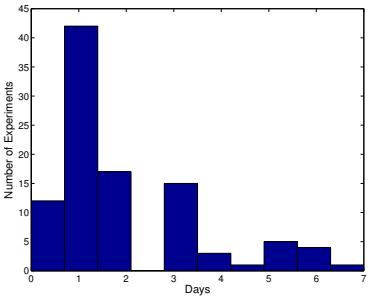

Fig. 3: Distribution of median value of time lag results in each experiment

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. Since the congressional record is much

larger, we subsample and repeat the experiment many times to get a distribution of time lags.

In each of these bootstrapped samples, there is a median time lag between the first appearance of a phrase in the congressional record and its first appearance in the media dataset. Figure 3 shows the distribution of these medians. The median is never negative, and is on average 2 days, showing a definite tendency for phrases to travel from the congressional record to the media rather than the other way round. The entire distribution also shows a slight bias towards the media picking up on congressional topics of discussion after the fact. These results help to explain the relative success of domain adaptation from the congressional record to the media dataset.

4 Conclusion

Text analytics is becoming a central methodological tool in analyzing political communication in many different contexts. It is obviously very valuable to have a good way of measuring political ideology based on text. Our work sounds a cautionary note in this regard by demonstrating the difficulty of classifying political text across different contexts. 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 being conservative or liberal, the concepts are distinct enough across datasets (even in just the US political context!) that generalization is extremely difficult. We note that, while we have presented our results in the context of classification, we get identical results when using measures of political ideology on a real-valued spectrum (the standard DW-Nominate score [26]) as the target of a regression task (this is only feasible for the congressional record, since the scores of congresspeople can be obtained as a function of their voting record). Our results demonstrate the need for extreme caution in the application of machine learning techniques to classifying political ideologies, especially when such efforts are made across domains.

References

1. Ahmed, A., Xing, E.P.: Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In: Proc. EMNLP. pp. 1140–1150 (2010)
2. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239), 1130–1132 (2015)
3. Baum, M.A., Groeling, T.: New media and the polarization of American political discourse. Polit. Comm. 25(4), 345–365 (2008)
4. Brown, A.R.: Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. PS: Polit. Sci. & Politics 44(02), 339–343 (2011)
5. Budak, C., Goel, S., Rao, J.M.: Fair and balanced? Quantifying media bias through crowdsourced content analysis. Public Opin. Quarterly 80(S1), 250–271 (2016)
6. Chen, M., Weinberger, K.Q., Xu, Z., Sha, F.: Marginalized stacked denoising autoencoders for domain adaptation. In: Proc. ICML. pp. 767—-774 (2012)
7. Clinton, J., Jackman, S., Rivers, D.: The statistical analysis of roll call data. Am. Polit. Sci. Rev. 98(02), 355–370 (2004)

8. Cohen, R., Ruths, D.: Classifying political orientation on twitter: It's not easy! In: Proc. ICWSM (2013)
9. Das, S., Lavoie, A., Magdon-Ismail, M.: Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. ACM Trans. on the Web 10(4), 24:1–24:25 (2016)
10. DellaVigna, S., Kaplan, E.: The Fox News effect: Media bias and voting. Q. J. Econ. 122(3), 1187–1234 (2007)
11. Entman, R.M.: How the media affect what people think: An information processing approach. J. Politics 51(02), 347–370 (1989)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR 17(59), 1–35 (2016)
13. Gentzkow, M., Shapiro, J.M.: What drives media slant? Evidence from U.S. daily newspapers. Econometrica 78(1), 35–71 (2010)
14. Gerrish, S., Blei, D.M.: Predicting legislative roll calls from text. In: Proc. ICML. pp. 489–496 (2011)
15. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proc. ICML. pp. 513–520 (2011)
16. Greenstein, S., Zhu, F.: Is Wikipedia biased? Am. Econ. Rev. 102(3), 343–348 (2012)
17. Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis pp. 1–31 (2013)
18. Groseclose, T., Milyo, J.: A measure of media bias. Q. J. Econ. 120(4), 1191–1237 (2005)
19. Gross, J., Acree, B., Sim, Y., Smith, N.A.: 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 (2013)
20. Ho, D.E., Quinn, K.M.: Measuring explicit political positions of media. Q. J. Polit. Sci. 3(4), 353–377 (2008)
21. Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political ideology detection using recursive neural networks. In: ACL. pp. 1113–1122 (2014)
22. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. KDD. pp. 497–506. ACM (2009)
23. Lin, W., Xing, E.P., Hauptmann, A.G.: A joint topic and perspective model for ideological discourse. In: Proc. ECML-PKDD. pp. 17–32 (2008)
24. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: Proc. ICLR (2013), http://arxiv.org/abs/1301.3781
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. JMLR 12, 2825–2830 (October 2011)
26. Poole, K.T., Rosenthal, H.: Congress: A political-economic history of roll call voting. Oxford University Press (1997)
27. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proc. EMNLP. pp. 1201–1211 (2012)
28. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proc. EMNLP. pp. 151–161 (2011)
29. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. EMNLP. pp. 1631–1642 (2013)
30. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proc. EMNLP. pp. 327–335 (2006)

Appendix

A Datasets

A.1 Congressional Record.

The U.S. Congressional Record preserves the activities of the House and Senate, including every debate, bill, and announcement. We use the party affiliation of the speaker (Democrat or Republican) as an indication of ideology (liberal or conservative). 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 excluded speech from independents, and from clerks and presiding officers.

A.2 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.

A.3 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 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. 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 4 shows the counts of articles in the liberal and conservative parts of each of the three datasets by year. Our datasets have the following properties that make them useful for political ideology learning and evaluation in the context of U.S. politics:

- The content is selected to be relevant to U.S. politics.
- The content can predictably be labeled as conservative or liberal by a somewhat knowledgeable human. While it is true that not all speeches by Democrats are liberal, and not all articles on Townhall conservative, since these are subjectively defined, this is nevertheless as clean a delineation as we can hope for.
- The creation times of items in the three datasets have substantial overlap;

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 4: Article distributions by year in the three datasets. Democrat (CR), Salon, and RationalWiki are assumed to be liberal, while Republican (CR), Townhall, and Conservapedia are assumed to be conservative.

A.4 Wikipedia

We also motivate our task by attempting to classify bias on Wikipedia, an important task [9]. Wikipedia is the largest encyclopedia project in the world and is widely used in both natural language processing and political science studies [4, 24]. Wikipedia is considered to have become nonpartisan as many users have contributed to political entries [16]. We focus on edits made by admins on political topics in Wikipedia. We download the English Wikipedia dump from March 4, 2015. To focus on US politics, we extract all articles (with full edit history) that belong to WikiProject United States[3] and satisfy the same political keywords requirement that we use for RationalWiki, yielding 4659 articles in total. We then collect all edits added or subtracted by each active Wikipedia admin.

B Details of Experimental Methodology

B.1 Cross-domain consistency

For the Congressional Record and Salon/Townhall datasets, we use data from 2005 to 2014. For the RationalWiki/Conservapedia datasets, we use the data from 2014 as capturing a recent snapshot. For this dataset only we use feature hashing to project the bigram features into a lower dimensional non-sparse feature space. We set the dimension of the hashed vector $n\_features = 20000$, $ngram\_range = (2, 2)$, and $decode\_error = ignore$. We use a so-called "balanced" logistic regression classifier to deal with the problem of class imbalance. All other parameters are the defaults in the scikit-learn package for both feature hashing vectorizer and logistic regression classifier.

B.2 Consistency across time

We use the TF-IDFLR method for this experiment. For the vectorizer, we set $min\_df = 5$, $ngram\_range = (2, 2)$ and $decode\_error = ignore$. For logistic regression classifier, we set $class\_weight = balanced$ to re-weight training samples. Other parameters are set to the default values in the scikit-learn package.

[3] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_United_States`

B.3 Domain adaptation

The linear classifier is the TF-IDFLR method described above. The RAE algorithm trains embeddings using sentences subsampled from the data in order to balance conservative and liberal training 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. We use five-fold cross validation when the training and testing sets are the same.