Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems[*]

Andrew Peterson[†] Arthur Spirling[‡]

Abstract

Measuring the polarization of legislators and parties is a key step in understanding how politics develops over time. But in parliamentary systems—where ideological positions estimated from roll calls may not be informative—producing valid estimates is extremely challenging. We suggest a new measurement strategy, that makes innovative use of the 'accuracy' of machine classifiers, i.e. the number of correct predictions made as a proportion of all predictions. In our case, the 'labels' are the party identifications of the members of parliament, predicted from their speeches, along with some information on debate subjects. Intuitively, when the learner is able to discriminate members in the two main Westminster parties well, we claim we are in a period of 'high' polarization. By contrast, when the classifier has low accuracy—and makes a relatively large number of mistakes in terms of allocating members to parties based on the data—we argue parliament is in an era of 'low' polarization. This approach is fast and substantively valid, and we demonstrate its merits with simulations, and by comparing the estimates from 78 years of House of Commons speeches with qualitative and quantitative historical accounts of the same. As a headline finding, we note that contemporary British politics is approximately as polarized as it was in the mid-1960s—that is, in the middle of the 'post-war consensus'. More broadly, we show that the technical performance of supervised learning algorithms can be directly informative about substantive matters in social science.

Word count: 3136 (excluding abstract and Online Appendices)

[*] We are grateful to Niels Goet, Justin Grimmer and Ben Lauderdale for comments an earlier draft. Audiences at the European Political Science Association meeting and the American Political Science Association meeting provided helpful feedback. Comments from two anonymous referees and the editor at Political Analysis improved our manuscript considerably. Our replication materials for this paper may be found here: http://dx.doi.org/10.7910/DVN/YTPJ1N

[†] Postdoctoral Researcher, University of Geneva. andrew.peterson@unige.ch

[‡] Associate Professor of Politics and Data Science, New York University. arthur.spirling@nyu.edu

1 Motivation

Understanding how well a supervised algorithm classifies new ('out-of-sample') examples is vital for assessing its utility for a given task. Thus in political science, to verify that a learning approach works well for a given categorization problem, we might compare the labels assigned by a trained machine to those given by humans to news stories (e.g. D'Orazio et al., 2014) or blog posts (e.g. Hopkins and King, 2010). Relatedly, in seeking to understand what types of words typify elite ideological divisions in the United States, we might inspect the performance of a given model to verify that the textual features we identify do an adequate job of differentiating the senators of different parties (e.g. Diermeier et al., 2012). But, in this *Letter* we put supervised model performance to a very different end: we show that, though these measures are designed for technical evaluation, they can also tell us something important directly and substantively about politics. In particular, we demonstrate that machine learning 'accuracy' provides an informative measurement instrument for the degree of aggregate polarization in the UK House of Commons over time.

To define terms explicitly: in keeping with the Americanist literature (e.g. Barber and McCarty, 2015), we understand 'polarization' to mean the (average) difference between the positions of the two main parties who have held Prime Ministerial office in modern times.[1] That is, the Labour (left/liberal) and Conservative (right/conservative) parties. Our central logic is to conceive of Members of Parliament (MPs) from different parties as being more or less distinguishable over time, in terms of what they choose to say. How distinguishable they are in practice is determined by a set of machine learning algorithms. Put very crudely, after being trained on a portion of the speeches, the models are then required to predict the most likely 'label'—that is, party identity—of the speeches that remain. When the machine

[1] See Online Appendix *A* for more details on our philosophy here.

learning accuracy—in the technical sense—is low, Labour MPs cannot easily be told apart from Conservative MPs (at least in terms of their speech contents). We deduce then that we are in a world of relatively low polarization. By contrast, when accuracy is high, and the machine does well at discriminating between partisans based on their utterances—say, with regards to the topics they raise, or the way they express themselves—we are in a more polarized era. As we show, these techniques provide a fast and valid way to estimate aggregate polarization that accords with simulation evidence, the historical record, and other data sources.

Before describing our data and approach, we note in passing that, on the substantive side, Britain's Westminster system is old and much imitated (Rhodes and Weller, 2005) and that its purported polarization has received a great deal of qualitative attention (e.g. Seldon, 1994). On the quantitative side, unlike in the Americanist literature (e.g. Barber and McCarty, 2015), we cannot generally use roll calls to infer relative partisan difference because (a) parties tend to vote extremely cohesively in the UK and (b) even when they don't, it can be difficult to interpret deviations substantively (Spirling and McLean, 2007). Scholars have measured ideology by surveying members (e.g. Kam, 2009) or by modeling networks of co-signing of initiatives (e.g. Kellermann, 2012), but data availability problems make this difficult to extend outside of the modern period. There are methods of positioning parties (e.g. Slapin and Proksch, 2008) and members (e.g. Lauderdale and Herzog, 2016), but these do not measure polarization explicitly, and tend to be computational intensive for large data sets.

2 Data: 3.5 million speeches over 78 years

Our data is essentially the entirety of the *Hansard* record of British parliamentary debates from 1935–2013.[2] This data has been extensively cleaned and matched with (disambiguated) meta-data on member names, ministerial roles and party identifications.[3] We study the two 'main' parties, Labour and Conservative, who controlled Prime Ministerial office for the entire period. We are working with a total of 3,573,778 speeches over 78 sessions, and we drop any speech with fewer than 40 characters, or which contain no words. The data shows balance between the parties, and encouraging consistency over time.[4]

We assume that the standard 'bag of words' vector space model is appropriate for the texts, with some preprocessing: we treat each speech as a series of token-specific (i.e. word-specific) frequencies that have been normalized by their maximum absolute value, which allows us to maintain the data in sparse format. We make no attempt to retain word order. We begin by fixing a vocabulary across all sessions[5] in which we drop any word that does not appear in 200 speeches in the entire dataset. This leaves 24,726 words. We do not stem or stop, or otherwise limit tokens, relying instead on the regularization process to drop unimportant terms.

3 Machine Learning Polarization

As the intuition above makes clear, our machine learning approach aims to capture the extent to which it is possible to distinguish between members of the two parties based on their speeches. We do this by using various supervised algorithms to predict the party

[2] Our replication materials for this paper may be found here: http://dx.doi.org/10.7910/DVN/YTPJ1N

[3] We obtained xml copies of the records from Kaspar Beelen. See Rheault et al. (2016) for details.

[4] See Online Appendix *B*

[5] One advantage of fixing the vocabulary is that it ensures that our measure is not subject to the bias identified by Gentzkow, Shapiro and Taddy (2016). See Online Appendix *C* for more details.

affiliation of the speaker of each speech in a legislative session. That is, we have labeled data—Conservative or Labour—and we seek to 'learn' the relationship between the speech information and the labels. We can report both an overall accuracy for our classifier, and provide estimates for any given MP in terms of their probability of being in one of the two (Conservative, Labour) classes, given their speeches and the relationships observed in the data.

As usual with machine learning approaches, we seek to balance strong predictive power against other concerns such as simplicity, reproducibility, overfitting, and computational time (see Hastie, Tibshirani and Friedman, 2009, for discussion of these issues). We chose four algorithms that embody all these features to varying extents. These are:

- the perceptron algorithm (see Freund and Schapire, 1999), a simple linear classifier with no regularization penalty and a fixed learning rate. This is trained by stochastic gradient descent, and is thus a special case of the second classifier:

- a stochastic gradient descent (SGD) classifier, which updates parameters on batches of randomly selected subsets of the data (for an overview see Bottou, 2004).

- the 'passive aggressive' classifier with hinge-loss, which updates parameters by seeking in each step a hyperplane that is close to the existing solution but which aggressively modifies parameters in order to correctly classify at least one additional example (Crammer et al., 2006).

- logistic regression with an L2 penalty, with regulation parameter $C = \frac{1000}{\#\text{ training speeches}} \approx 0.2$, fit using stochastic average gradient descent (see Schmidt, Roux and Bach, 2013).

Within each legislative session, we run all four algorithms, then select the algorithm with the highest accuracy as the representative of that session. All four algorithms are implemented using `Scikit-Learn` (Pedregosa et al., 2011) in the Python language. For each classifier we also average the accuracy over a stratified 10-fold cross-validation. Though different in nature, the algorithms perform extremely similarly, on average, which suggests there is little model dependence to our findings (see Online Appendix D).

Different legislative sessions have different numbers of members and speeches by one party or the other. We use class (party) weights inversely proportional to the class (party) frequencies, i.e. $\frac{n}{2 \cdot n_p}$, where n is the total number of speeches and n_p is the number of speeches by members of that party. That is, we essentially weight up the speeches of the less commonly observed party in a given session for the purpose of training the classifiers.

For every *speech*, with no loss of generality, we produce an estimated probability that it was given by a Conservative member (the probability that is was given by a Labour member is simply one minus that estimate). The probability that a given *member* is a Conservative is then the mean of the probabilities of all their speeches. In the usual way, we allocate (predict) a discrete class label of 'Conservative' to all MPs with (mean speech) probability $\geq \frac{1}{2}$, and 'Labour' otherwise. For a set of MPs in a session, the *accuracy* of the classifier is

$$\frac{|\text{true positives}| + |\text{true negatives}|}{|\text{true positives}| + |\text{true negatives}| + |\text{false positives}| + |\text{false negatives}|}$$

where the terms are as described in Table 1, and $|\cdot|$ indicates the raw number of each quantity.

We note that estimation of the models is fast (less than one second per classifier per session) so that even with the 10-fold cross-validation more time is spent loading and preparing the data than running the algorithm. Ignoring this data preparation time, fitting our classifiers and predicting labels for all speeches required a total of 22.6 minutes.

Term	True Label	Machine Assigned Label
true positive	Conservative	Conservative
true negatives	Labour	Labour
false positive	Labour	Conservative
false negative	Conservative	Labour

Table 1: Definition of terms for accuracy calculation

In terms of related literature, our work is similar in spirit to recent efforts from Gentzkow, Shapiro and Taddy (2016). Those authors also provide a method for estimating polarization from speeches. Importantly, it avoids bias that can arise from sampling error when aggregating differences in high-dimensional count data. That technique is generative and model-based, which may well be preferable for some researchers. In contrast to their "highly parametric" approach, ours is nonparametric and can be quickly scaled to millions or billions of documents (see e.g. Chen and Guestrin, 2016). By contrast, Gentzkow, Shapiro and Taddy (2016) obtain scalability by using a Poisson approximation to the relevant likelihood.

Before moving to the results, we make two points about the scope of our work here. First, as with roll call based discussions of polarization, our measure can tell us only about the *relative* level of polarization at one time as against another. Consequently, our aim is not high predictive accuracy *per se* but rather predictive consistency: i.e. a maintained assumption is that variations in accuracy from one time period to another are indeed a result of substantive differences in speeches and not an artifact of data collection problems or the failure of the algorithm to identify the relevant features. Second, we used an ensemble method (gradient boosted trees) to verify the plausibility of this assumption. The idea is that while more computationally intensive and more difficult to interpret than our four options above, such a technique may achieve higher accuracy and thus enable us to diagnose whether the variation we see in performance below is simply due to the idiosyncratic choices of algorithms we made and the way they handle the data they receive. As expected, the ensemble method achieved a significant increase in accuracy (mean of 0.80 instead of 0.74). Critically, however, the new measure produces the same over-time variation and thus suggests our approach reliably captures relative differences in polarization over time rather than statistical artifacts (see Online Appendix E for discussion).

4 Results and Validation

Does this method work for measuring polarization in practice? We now turn to a series of validations suggesting it does. We begin with simulations—where we know the truth by construction—and seek to verify our technique recovers parameters appropriately.

4.1 Validation I: Simulation Evidence

First we want to show that *if* the parties differ systematically in terms of the tokens they use, our approach separates them as an increasing function of that difference in vocabulary.

We model speech as follows. There are three types of words: ‘left’ and ‘right’ which have no overlap, and ‘noise’ words which have no relationship to partisanship. For a fixed degree of a speech which is noise, for the rest of the speech token slots, a Conservative (Labour) member chooses a ‘right’ (‘left’ in the Labour case) word with probability $a \geq \frac{1}{2}$ and a ‘left’ (‘right’) word with probability $1 - a$. We denote a the ‘separation’ parameter, and as it approaches 1, polarization is increasing. At $a = 1$, members use completely disjoint partisan vocabularies, and their speeches overlap only in terms of noise words. A ‘parliament’ is 600 members, half from each party, with each giving one speech of 100 words selected as discussed. We perform a TFIDF weighting of the relevant matrix, apply the learner(s), and output a predicted probability that each speech/member is Conservative.

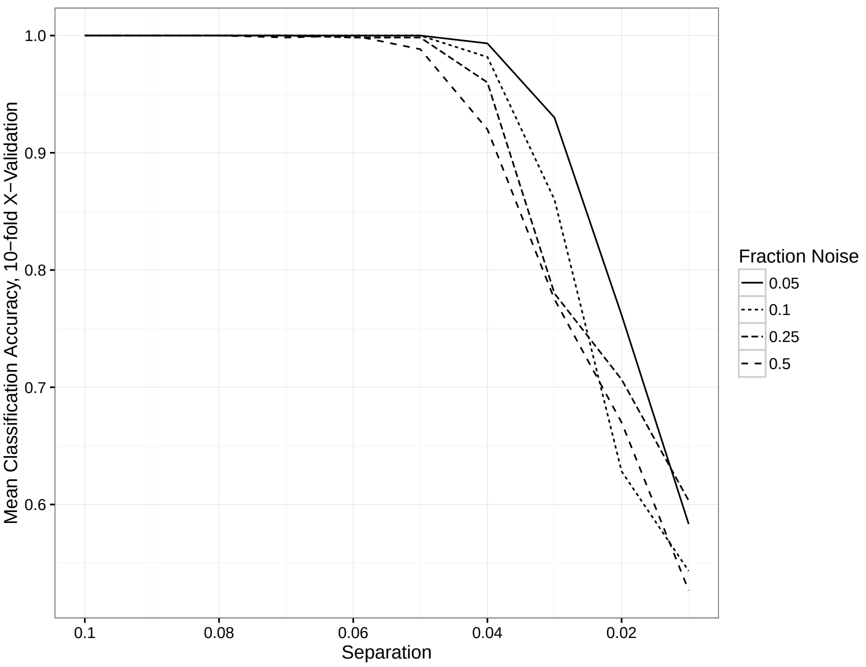

Figure 1: Classification Accuracy (y-axis) for Different Levels of Separation (x-axis) at different levels of noise.

As hoped, as a increases for a fixed degree of noise $(0.05, 0.1, 0.25, 0.5)$, we see from Figure 1 that accuracy—i.e. polarization—increases. There, the x-axis represents values of a. When the separation is sufficiently large at these noise levels ($a \gtrsim 0.06$, though these magnitudes are not directly interpretable), the classification rate (on the y-axis) is perfect (1.0). As the two parties become more similar in their word choices, the classification accuracy declines until the algorithm is doing no better than chance (at separation ≈ 0.01).

Second, we want to explore the relationship between our measure of polarization and noise. It is conceivably the case that as noise (i.e. the frequency of non-partisan terms) increases—perhaps due to new topics or parliamentary procedures that arise—our method will suggest the parties are converging, whereas they remain as different at their core as they were previously. Figure 2 shows the (bimodal, Labour-Conservative) density of estimates of the predicted probability of being Conservative for each of the 600 speeches, while fixing the

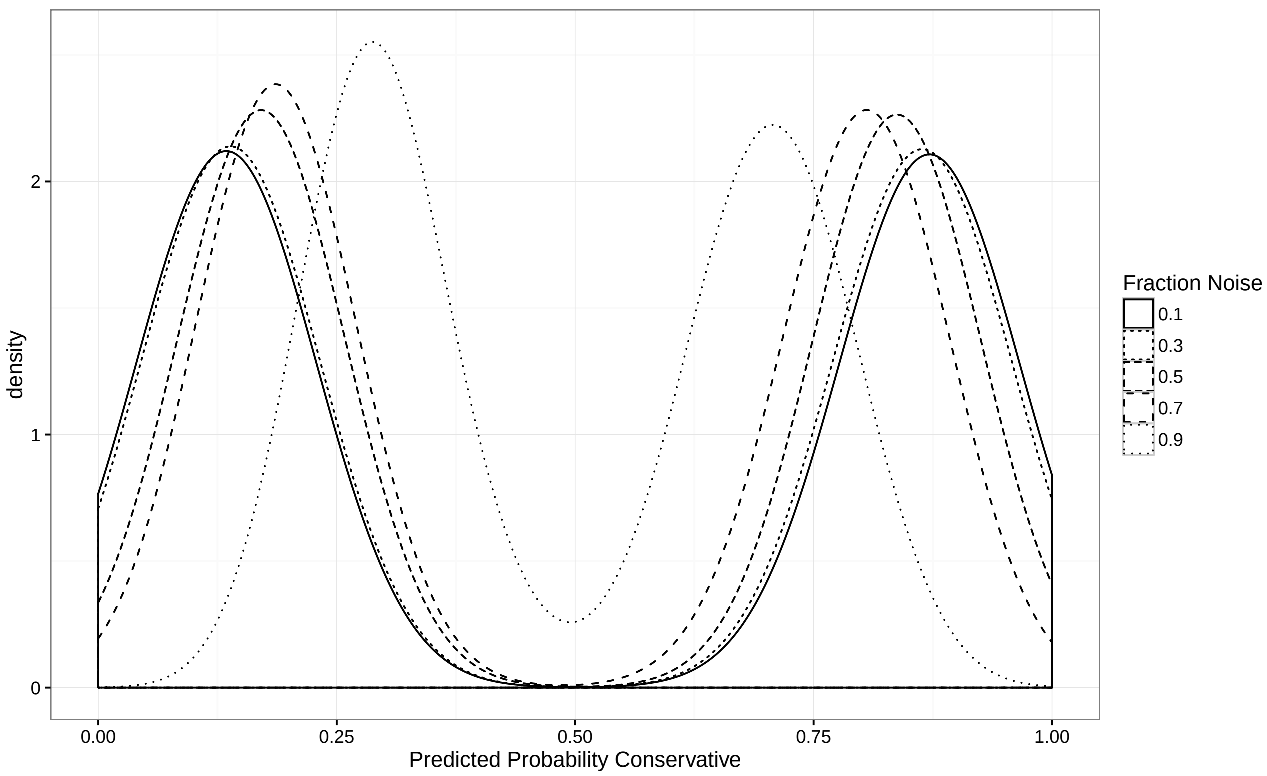

Figure 2: Density Plot of Predicted Probability Conservative For Different Levels of Noise. Note that as the fraction of noise in the data generating process increases, the mean positions of the parties are forced closer together.

difference in the two parties (at separation = 0.1). We allow for the fraction of the words that are noise to vary from 0 to 0.9. When the words are less than 60% noise, there is little artificial change in polarization as a function of noise: the parties, on average, stay close to the extremes. But it is also true that as noise increases, the parties falsely appear more similar. From other experiments we did,[6] it became apparent that in such a high noise situation, the variance with which *each member* is estimated is also higher. This suggests that we can identify the difference between true ideological moderation and the presence of noise by looking for changes in the precision with which members' positions are estimated over time. We return to this point below.

[6] See Online Appendix *F*.

4.2 Validation II: Qualitative Historical Record

We plot our session accuracy results in Figure 3, and it strongly accords with our priors and those of others for the period (Addison, 1994; Seldon, 1994; Fraser, 2000). In the 1930s, polarization drops rapidly, reaching a nadir in the years of the Second World War. This makes sense given the (Churchill led) coalition government of that time. Soon after, when elections begin in earnest with the 1945 Labour landslide, polarization ticks up. It then enters a long period of approximate stasis—the 'post-war consensus' (Kavanagh and Morris, 1994)—between circa 1945 and circa 1979, with small movements around the mean, though it is gradually sloping upwards. From the first session of 1979, i.e. the session in which Margaret Thatcher assumed the premiership, polarization jumps and reaches its zenith around the session corresponding to 1987. It then falls, gradually at first and then more quickly, as Tony Blair becomes leader of Labour after 1994. By the sessions around 2001, polarization is falling sharply, with the end of Gordon Brown's government and the beginning of the Conservative-Liberal Democrat coalition marking a further decline. The dark vertical [green] lines represent structural breaks, in the sense of Bai and Perron (2003) (as implemented by Zeileis et al. (2002)). These provide more formal evidence of our validation claims, with change points in September 1948, November 1978 and June 2001. We note in passing that, by our estimates, polarization in the contemporary House of Commons is on a par with that of the mid-1960s.

Figure 4 presents the mean variance in speaker estimates for the time period under study. Importantly, it not noticeably higher during claimed periods of consensus (i.e. post-war). This is good news, and implies that—per Section 4.1—the measure does indeed capture a change in ideological polarization rather than an artifact of any changing noisiness of speeches.

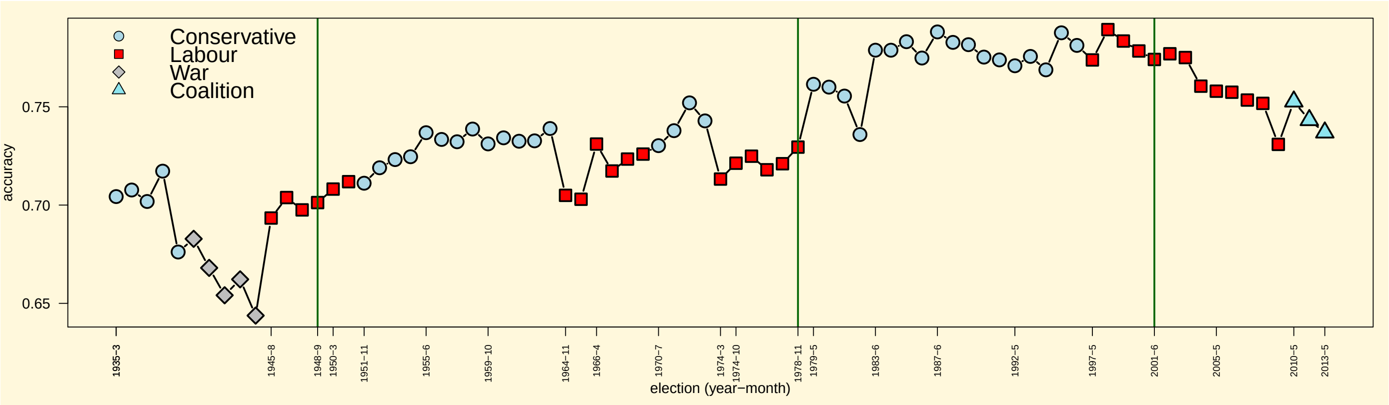

Figure 3: Estimates of parliamentary polarization, by session. Election dates mark x-axis. Estimated change points are [green] vertical lines.

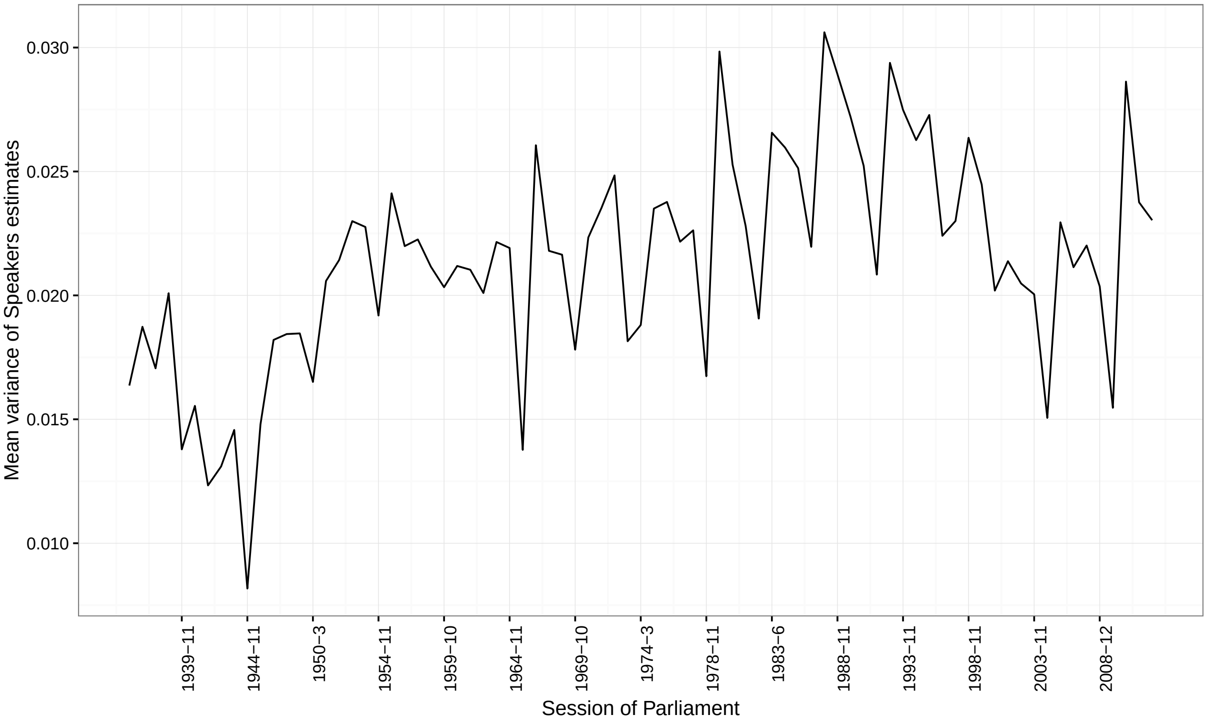

Figure 4: Mean Variance by Session

4.3 Validation III: Quantitative Historical Record

We can also compare our accuracy results to more quantitative evidence. In Figure 5 we plot the two main UK parties in terms of their manifesto 'RILE' scores (a measure of where they lie in some overall sense on the standard left-right spectrum) as provided by the Manifesto Project (Lehmann et al., 2016; Volkens et al., 2016) for the post-1945 period. The individual points refer to parties in different years (with higher scores implying positions are more right wing), while the solid line is the (absolute) difference between the parties. The broken line is a lowess of the same. When these lines are relatively high, the parties are more polarized (literally more different). When they fall, the parties are closer together.

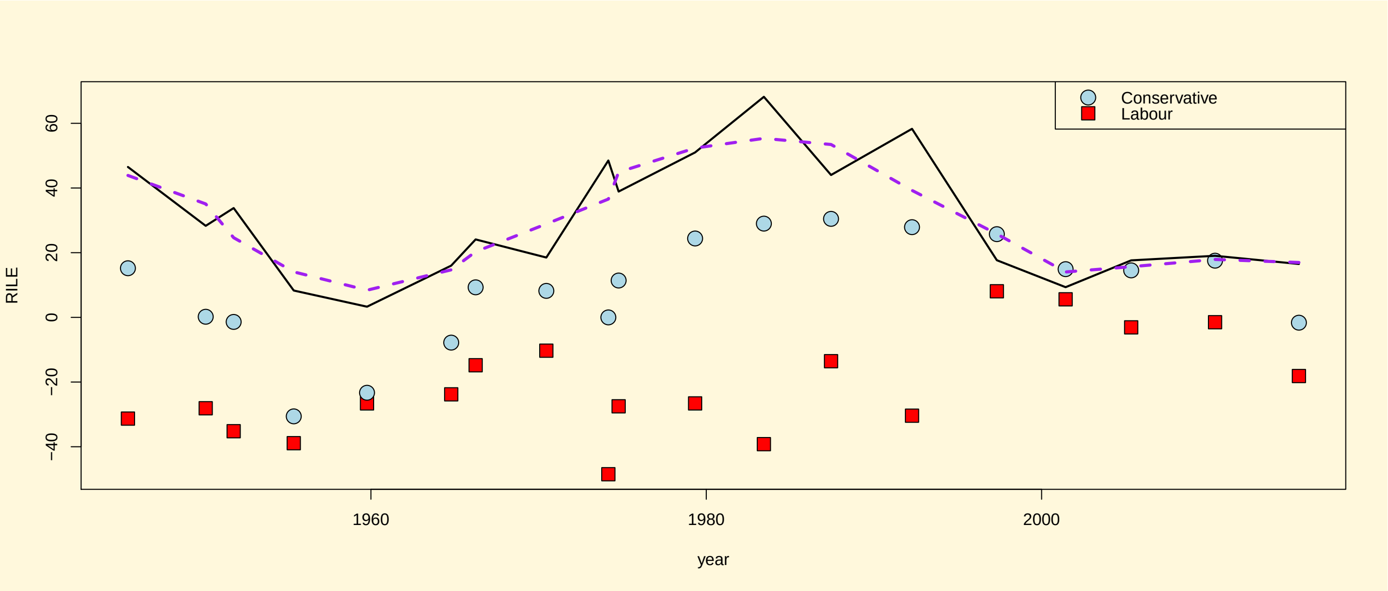

Figure 5: Left/right (RILE) scores from the Manifesto Project. Higher scores correspond to more right wing policies. Lines are difference between the parties (solid) and lowess (broken) of the same.

Of course, manifestos are written prior to a parliament being formed, and there are many reasons to believe the polarization we see in electoral promises may not show up in identical magnitudes in a legislature. Comfortingly though, we see the same broad pattern as in Figure

3: polarization is relatively low after the war, reaching a peak in the Thatcher years, before entering secular decline again. Comparing the manifesto dates to the closest parliamentary session, we note a reasonable positive correlation of approximately 0.16.

5 Discussion

We argued that the performance of a classifier can be used to measure aggregate polarization in the UK, and that the estimates from this process accord with—and extend—other quantitative and qualitative evidence.[7] This approach is fast and replicable. From the simulation evidence, we strongly suspect it can be ported to other domains where traditional instruments, like roll calls, are either unavailable or uninformative. Obviously, there will be some limits: unsurprisingly, we anticipate that it will work best when parties that are relatively far apart on a given latent dimension do, indeed, use different vocabularies *when discussing the same issue*. This latter caveat is important: claims about polarization make most sense when parties (or people) have different perspectives on the same topics; that is, when they are not simply raising (possibly orthogonal) subjects of interest which have implicitly different word frequencies. So, institutional settings, where debate is free-flowing—in the sense that different 'sides' can use different vocabularies—but 'on-topic' are ideal. These might include parliaments working through a legislative agenda, committees working through a meeting schedule and courts discussing specific matters of law. Note that these institutional practices ought to be consistent: we expect our approach to perform poorly if there are changes to vocabulary forced on one 'side' but not the other. In general, inspecting the terms which discriminate between parties is helpful for knowing which situation pertains.[8]

[7] This includes roll call clustering studies for the UK: see Online Appendix *G* for a discussion, along with advice on validation in other contexts.

[8] We give more advice for practitioners in Online Appendix *H*.

Within the Westminster system, extending the central logic to more than two parties should be straightforward although some thought is required in terms of the direct interpretation of the output in that case. Ultimately, our approach is based on estimates of speeches and the individual MPs that made them: future work might make direct use of those estimates after careful validation.

References

Addison, Paul. 1994. *The Road to 1945: British Politics and the Second World War*. London: Pimlico.

Bai, Jushan and Pierre Perron. 2003. "Computation and Analysis of Multiple Structural Change Models." *Journal of Applied Econometrics* 18:1–22.

Barber, Michael and Nolan McCarty. 2015. Causes and Consequences of Polarization. In *Solutions to Polarization in America*, ed. Nathaniel Persily. Cambridge: Cambridge University Press pp. 15–59.

Bottou, Léon. 2004. Stochastic learning. In *Advanced lectures on machine learning*. Springer pp. 146–168.

Chen, Tianqi and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM pp. 785–794.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz and Yoram Singer. 2006. "Online Passive-Aggressive Algorithms." *Journal of Machine Learning Research* 7(1):551–585.

Diermeier, Daniel, Jean-Franois Godbout, Bei Yu and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42:31–55.

D'Orazio, Vito, Steven Landis, Glenn Palmer and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22(2).

Fraser, Duncan. 2000. "The Postwar Consensus: A Debate Not Long Enough." *Parliamentary Affairs* 53(2):347–362.

Freund, Yoav and Robert E. Schapire. 1999. "Large Margin Classification Using the Perceptron Algorithm." *Machine Learning* 37(3):277–296.

Gentzkow, Matthew, Jesse M Shapiro and Matt Taddy. 2016. "Measuring Polarization in High-dimensional Data: Method and Application to Congressional Speech." *NBER Working Paper* .
URL: *http://www.nber.org/papers/w22423*

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hopkins, Daniel and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Kam, Christopher J. 2009. *Party Discipline and Parliamentary Politics*. Cambridge: Cambridge University Press.

Kavanagh, Dennis and Peter Morris. 1994. *Consensus Politics from Attlee to Major*. Hoboken: Wiley Blackwell.

Kellermann, Michael. 2012. "Estimating Ideal Points in the British House of Commons Using Early Day Motions." *American Journal of Political Science* 56(3):757–771.

Lauderdale, Benjamin and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24(2):1–21.

Lehmann, Pola, Theres Matthieß, Nicolas Merz, Sven Regel and Annika Werner. 2016. "Manifesto Corpus.". Version: 2016-6.

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

Rheault, L, Beelen K, Cochrane C and Hirst G. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLOS ONE* 11(12).

Rhodes, Rod and Patrick Weller. 2005. Westminster Transplanted and Westminster Implanted: Exploring Political Change. In *Westminster Legacies: Democracy and Responsible Government in Asia and the Pacific*, ed. Haig Patapan, John Wanna and Patrick Weller. University of New South Wales: University of New South Wales Press.

Schmidt, Mark, Nicolas Le Roux and Francis Bach. 2013. "Minimizing finite sums with the stochastic average gradient." *arXiv preprint arXiv:1309.2388* .
URL: *https://arxiv.org/abs/1309.2388*

Seldon, Anthony. 1994. "The Consensus Debate." *Parliamentary Affairs* 47(4):501–514.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52.

Spirling, Arthur and Iain McLean. 2007. "UK OC OK?" *Political Analysis* 15(1):85–96.

Volkens, Andrea, Pola Lehmann, Matthieß Theres, Nicolas Merz and Sven Regel. 2016. "The Manifesto Data Collection.". 2016b.

Zeileis, Achim, Friedrich Leisch, Kurt Hornik and Christian Kleiber. 2002. "strucchange: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software* 7(2):1–38.