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

Andrew Peterson[1] and Arthur Spirling[2]

[1] *Postdoctoral Researcher, University of Geneva, Switzerland. Email: andrew.peterson@unige.ch*
[2] *Associate Professor of Politics and Data Science, New York University, USA. Email: arthur.spirling@nyu.edu*

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 "postwar consensus". More broadly, we show that the technical performance of supervised learning algorithms can be directly informative about substantive matters in social science.

Keywords: statistical analysis of texts, polarization, learning

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

Authors' note: We are grateful to Niels Goet, Justin Grimmer and Ben Lauderdale for comments on 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 are as described in Peterson (2017).

Political Analysis (2018)
vol. 26:120–128
DOI: 10.1017/pan.2017.39

Corresponding author
Andrew Peterson

Edited by
Justin Grimmer

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 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 do not, 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 to 2013.[2] This data has been extensively cleaned and matched with (disambiguated) metadata 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.

1 See Online Appendix A for more details on our philosophy here, found in the supplementary material.
2 Our replication materials are as described in Peterson (2017).
3 We obtained xml copies of the records from Kaspar Beelen. See Rheault *et al.* (2016) for details.
4 See Online Appendix B in the supplementary material.
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 in the supplementary material for more details.

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 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 in the supplementary material).

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 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 on 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.

Table 1. Definition of terms for accuracy calculation.

Term	True Label	Machine Assigned Label
True positives	Conservative	Conservative
True negatives	Labour	Labour
False positives	Labour	Conservative
False negatives	Conservative	Labour

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 overtime variation and thus suggests our approach reliably captures relative differences in polarization over time rather than statistical artifacts (see Online Appendix E in the supplementary material 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.

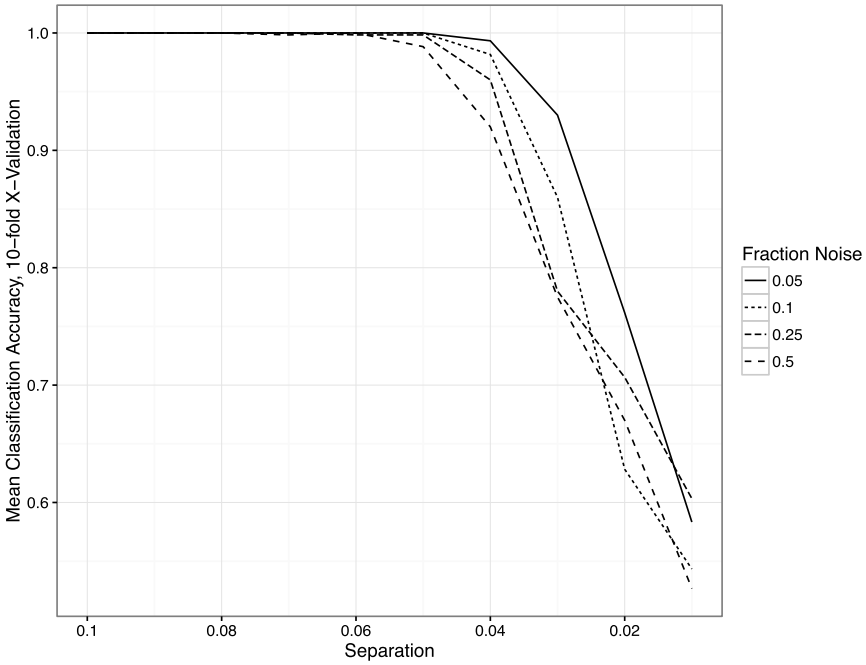

Figure 1. Classification accuracy (y-axis) for different levels of separation (x-axis) at different levels of noise.

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 $\approx$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 nonpartisan 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 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.

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

6 See Online Appendix F in the supplementary material.

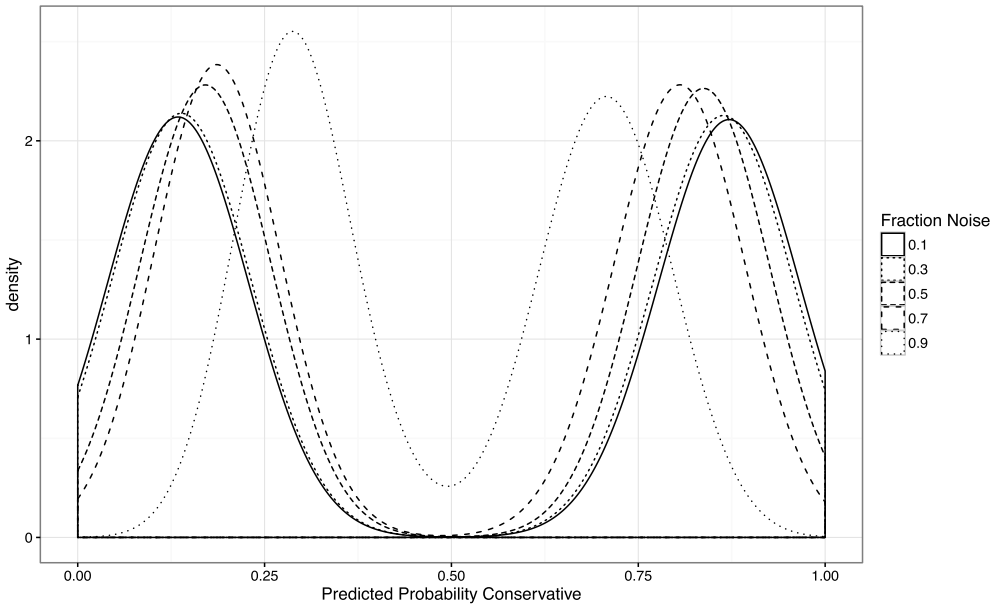

Figure 2. Density plot of predicted probability conservative for different levels of noise. Note that as the fraction of noise in the data generating process increases, the mean positions of the parties are forced closer together.

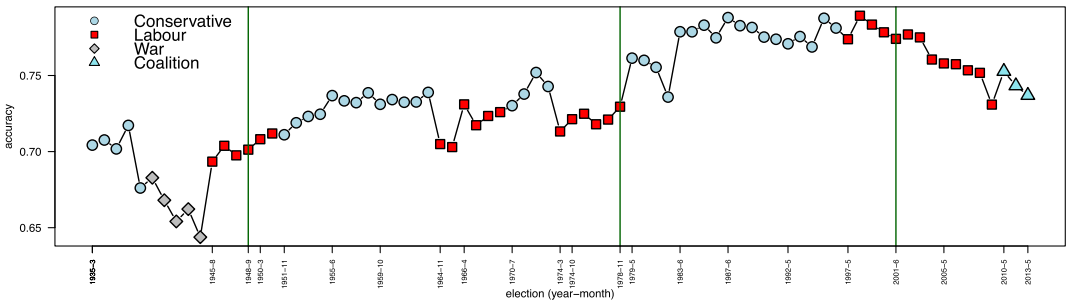

Figure 3. Estimates of parliamentary polarization, by session. Election dates mark *x*-axis. Estimated change points are [green] vertical lines.

with the 1945 Labour landslide, polarization ticks up. It then enters a long period of approximate stasis—the "postwar 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 is not noticeably higher during claimed periods of consensus (i.e., postwar). This

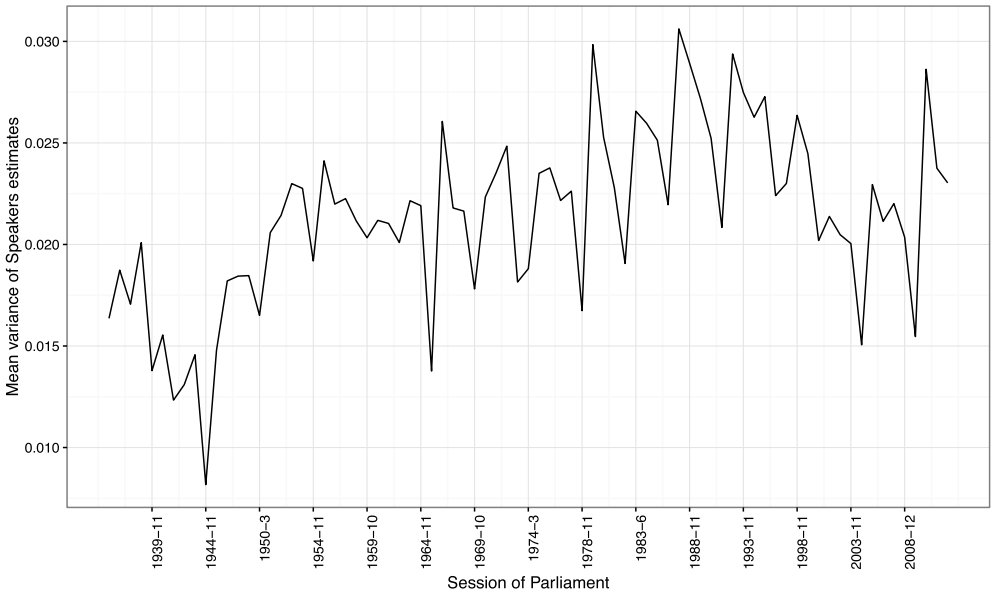

Figure 4. Mean variance by session.

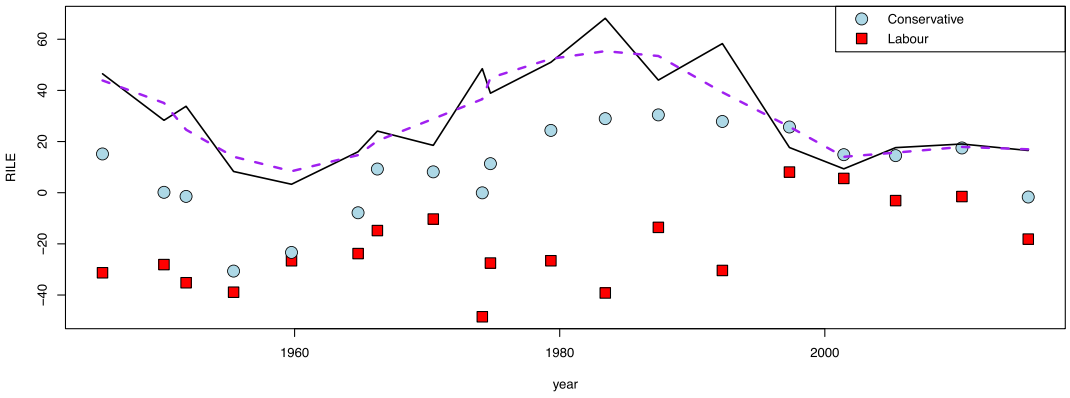

Figure 5. Left/right (RILE) scores from the Manifesto Project. Higher scores correspond to more right wing policies. Lines are difference between the parties (solid) and lowess (broken) of the same.

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.

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.

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]

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.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2017.39.

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. Persily Nathaniel. Cambridge: Cambridge University Press, pp. 15–59.

Bottou, Léon. 2004. Stochastic learning. In *Advanced lectures on machine learning: ML summer schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, revised lectures*, ed. O. Bousquet, U. von Luxburg, and G. Rätsch. Berlin, Heidelberg: Springer, pp. 146–168.

Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *KDD '16 proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: 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):224–242.

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

8 We give more advice for practitioners in Online Appendix H in the supplementary material.

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. 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, and M. Blondel et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.
Peterson, Andrew Jerel. 2017. Replication data for: Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. Harvard Dataverse, UNF:6:iRJ1F7aydu3LemeJ2gjS1A==. doi:10.7910/DVN/YTPJ1N.
Rheault, L., K. Beelen, C. Cochrane, and G. Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE* 11(12). URL: https://doi.org/10.1371/journal.pone.0168843.
Rhodes, Rod, and Weller Patrick. 2005. Westminster transplanted and westminster implanted: Exploring political change. In *Westminster legacies: Democracy and responsible government in Asia and the Pacific*, ed. Patapan Haig, 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. Preprint arXiv: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(3):705–722.
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. Manifesto Project (MRG/CMP/MARPOR). Version 2016b. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). URL: https://doi.org/10.25522/manifesto.mpds.2017a.
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.

ONLINE APPENDIX
Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems

Andrew Peterson* Arthur Spirling†

*Postdoctoral Researcher, University of Geneva. andrew.peterson@unige.ch
†Associate Professor of Politics and Data Science, New York University. arthur.spirling@nyu.edu

Online Appendix A Clarifying 'Polarization'

It is helpful to elucidate the difference between our measure of polarization and the underlying concept in politics that we believe it connotes. As discussed, 'polarization' is about discrimination: that is, when it is generally easier to distinguish the statements of one party from another, we consider the world more polarized. Substantively, we think of UK politics as existing on an essentially unidimensional line, from 'far left' (typically associated with the Labour party) to 'far right' (typically associated with the Conservative party) for the period under study. One could think of positions on that line as being weighted combinations of all or some of the salient policy issues of the day. Crucially, we are agnostic as to where on the line the parties are (on average) located at any particular time. That is, they might both be left of the period median, or right of the period median or somewhere else. What matters, instead, is how *different* they are from each other at that time: it is precisely this separation that gives rise to claims of polarization.

To fix ideas, consider politics immediately after the Second World War, versus politics around the 2010 election. We would argue that both times are periods of low polarization, even though the parties were in very different places (on average) on the relevant continuum. In particular, after the war, both Conservative and Labour parties accepted a large role for the state in industry, high public spending, relatively high personal taxation rates etc. That is, both parties were 'left' of the median of the period as a whole, but close to one another nonetheless. Whereas, by 2010, both parties accepted the privatizations of the Thatcher years along with relatively low public spending as fixed aspects of the landscape. That is, both parties were 'right' of the median of the period as a whole, but close to one another nonetheless. To reiterate, when we report that polarization for these periods is low, we mean that the parties were close to one another, not that they were centrist, or moderate, in some global sense.

Pushing beyond the data, our notion of polarization pertains to the difference one would expect to observe were one party in government replaced with the other. Of course, we measure everything at the speech level (it could simply be 'cheap talk'), but we would contend that the differences would be in terms of policy, too. This has particular resonance in Westminster systems because governments have large majorities and can generally enact the policies they espouse. Again, to fix ideas, we would argue that had Labour won the 1983 or 1987 general elections (a period of high polarization by our measure) voters would have seen very different policy enacted. By contrast, had Labour and Conservative parties changed places in government and opposition in the 1950s and 1960s (as they did), we would see relatively little change to policy as a whole (which is exactly what the historical consensus suggests).

Online Appendix B Temporal Stability of the Data

Our results are unlikely to be the spurious result of artificial long-term trends in how speeches are made in Parliament. In particular, while there is some variation from one session to another in the number and length of speeches given by members, there is no general trend that aligns with our findings about polarization. Consider first the number of speeches made by each member per session, presented in Figure 6. While there is some local cyclicality related to electoral periods (with a higher mean number of speeches given in 1979 when Thatcher was elected, for example), overall there is no detectable trend.

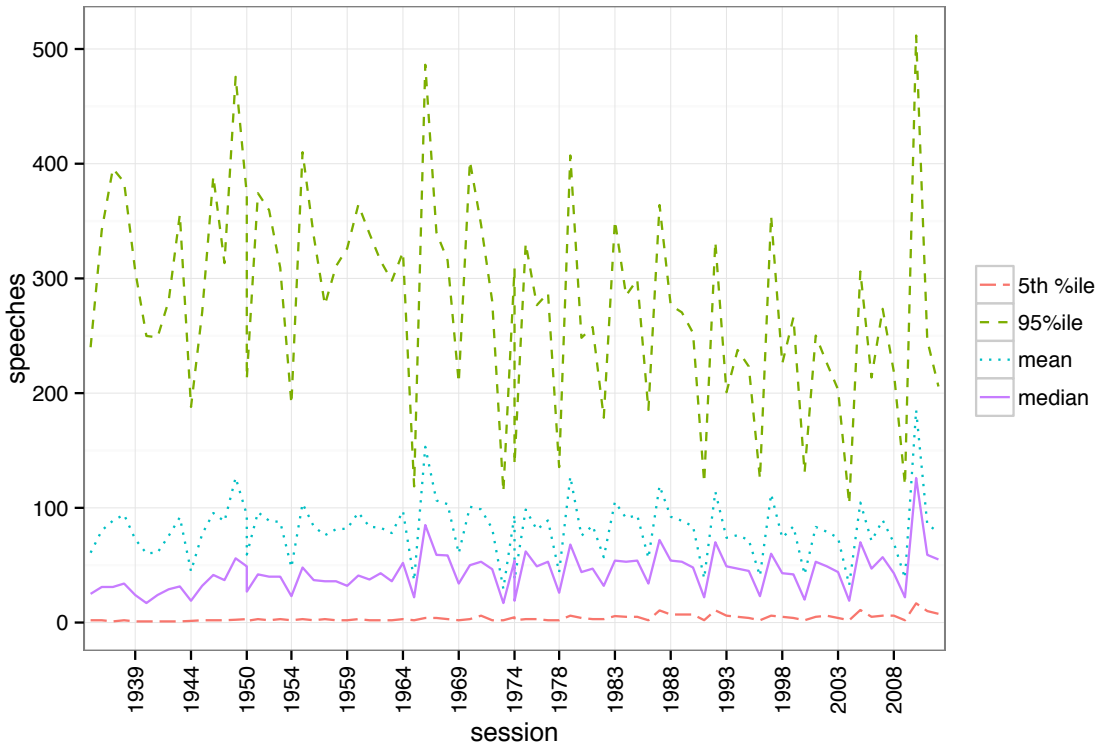

Figure 6: Number of Speeches By Member Per Session.

In addition to the number of speeches given, we might be concerned that there are differences in the length of speeches, which could reflect differences in cohorts or procedural roles played by different members. The evidence suggests this is not the case, however, as the mean length of speeches by different MPs remains constant throughout the period of our study. We present the mean and 5th, 50th, and 95th percentiles of the mean length of speeches in Figure 7. While there is a slight increase in the mean length in the post-war period and a slight decrease in recent years, this is minor and does not match the trends we

identify in our polarization measure.

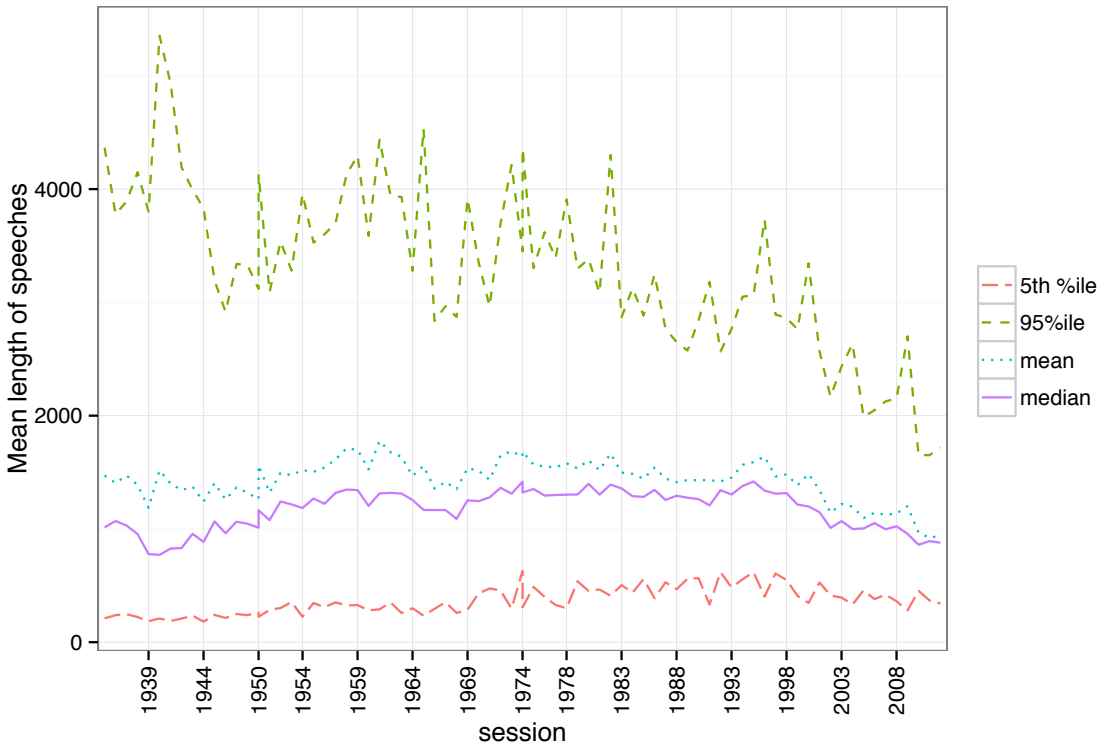

Figure 7: Mean Length of Speeches By Member Per Session.

As alluded to, the data is also remarkably well balanced in terms of partisan contributions, which is a testament to the dominance of the two 'big' British parties at this time. Thus, the Conservative party gave an average of 21,805 speeches per session, while the Labour party gave slightly more (23,432). Overall, each member gave an average of 1,128 speeches in their parliamentary career, with a mean of of 82 speeches in each session. Broken down by party, Tories gave an average of 83 speeches, while Labour members gave 81 speeches per session. The average Conservative speech was 1,023 characters, and for Labour speech it was 1,103 characters. This is comforting though, in any case, where there is asymmetry in representation we use class weights to ensure that the classifier will not increase accuracy by predicting the more common class.

Online Appendix C Measurement Concerns

C.1 Possible Bias from Size of Vocabulary changes

Gentzkow, Shapiro and Taddy (2016) show that two recent measures of polarization from

speech based on text (Gentzkow and Shapiro, 2010; Jensen et al., 2012) can be biased by changes in the size of the vocabulary. Such a critique could be of particular interest to our findings since they argue that the revised measure identifies significant polarization in recent years in the U.S. case. However, since we fix the vocabulary across all Parliamentary sessions, we have little reason to think this would affect our results. Their approach to demonstrating this, however, which involves comparing the results when party labels are randomly assigned by member, provides a way to examine whether our results may be the product of some other similar spurious relationship. In particular, we would be concerned if the trend line from the randomized labels closely tracks the trend of our measure (compare Gentzkow, et al, Figures 2, 3). This is not the case for our results, as is clear from comparing our results (in red) to those of 10 runs of randomized party labels (Figure 8). While there is some variation in the estimates generated from random labels, it does not match the overall pattern, and differs from them quite substantially at points, such as in suggesting high polarization during the World War II era.

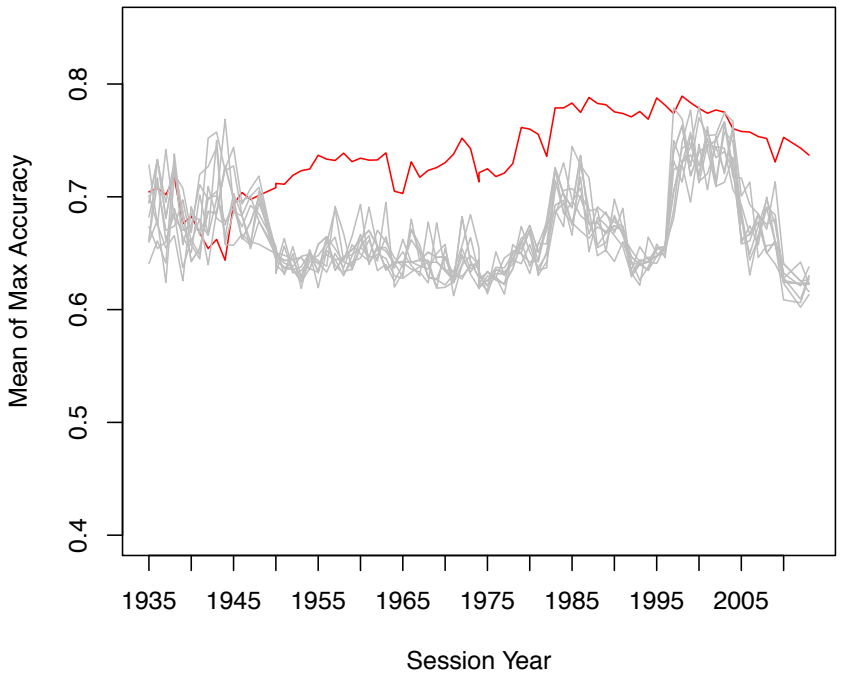

Figure 8: Estimates of parliamentary polarization, by session, by algorithm. The accuracy using real party labels (our polarization measure) is in red, while 10 runs with party labels randomized by speaker are presented in grey.

C.2 Uncertainty

Another measurement concern is that of the uncertainty of the estimates. Since our approach is not based on a generative model of text, we undertake a simple bootstrap, and we do this in two ways—one more conservative than the other. In particular, we resample from the set of speeches in each Parliamentary session 100 times, and we generate 10 folds for each as before. We then train and run the algorithms to calculate accuracy scores for each session.

The results are presented in Figures 9 and Figure 10. For the former figure, we take a 'naive' approach, and simply plot—for each point estimate from whatever the best performing algorithm was for that session—two standard errors on each side of the mean. Given that for each session we have between 15,000 and 104,000 speeches, these intervals are inevitably very narrow. In the second plot we provide a non-standard but, in this case, more conservative approach. Specifically, we calculate confidence intervals based on the 5th and 95th percentiles of the estimates for *each* of the four algorithms (rather than the highest performing) across the samples and folds.

Our main point is that the overall trend of the polarization measure is significant despite uncertainty over which texts are sampled—and this is true whichever way we perform the bootstrap.

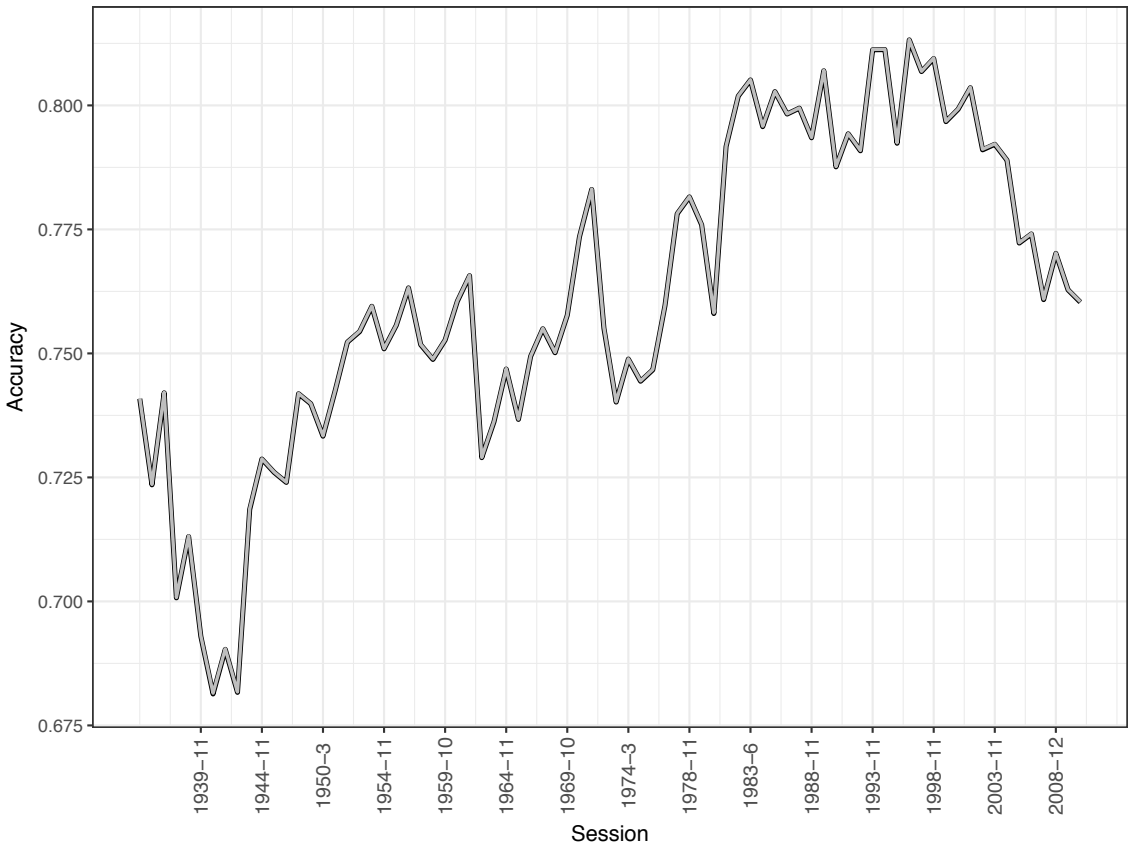

Figure 9: Polarization measure with bootstrap confidence interval, based on resampling texts within each session

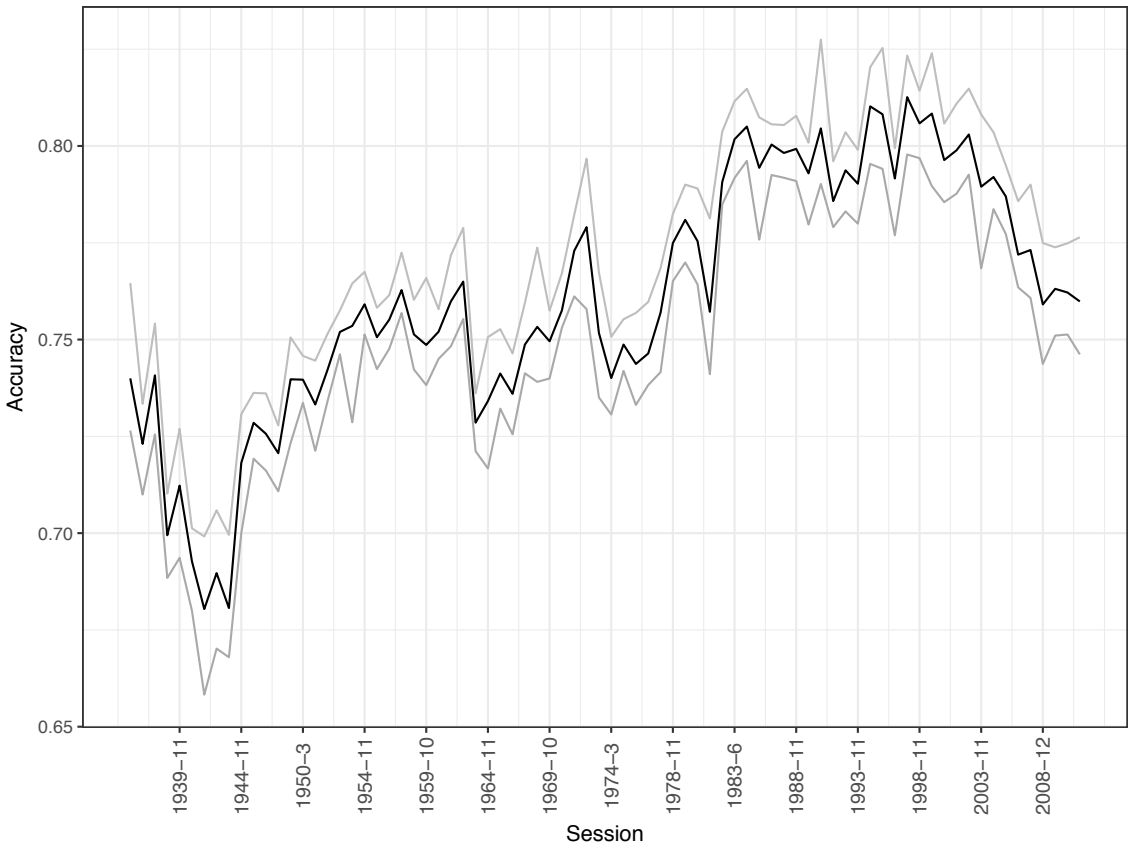

Figure 10: Polarization measure with percentile bootstrap confidence interval, based on all estimates from each of the four algorithms, while resampling texts within each session

Online Appendix D Machine Algorithms Produce Similar Results

Recall that we use four machine learning algorithms: perceptron and passive aggressive classifiers, a stochastic gradient descent classifier using a hinge-loss penalty and logistic regression using stochastic average gradient descent. When we inspect their mean accuracy rates over time, we see they perform almost identically. This is shown in Figure 11, where the lines each correspond to a different classifier and, importantly, are barely distinguishable from one another.

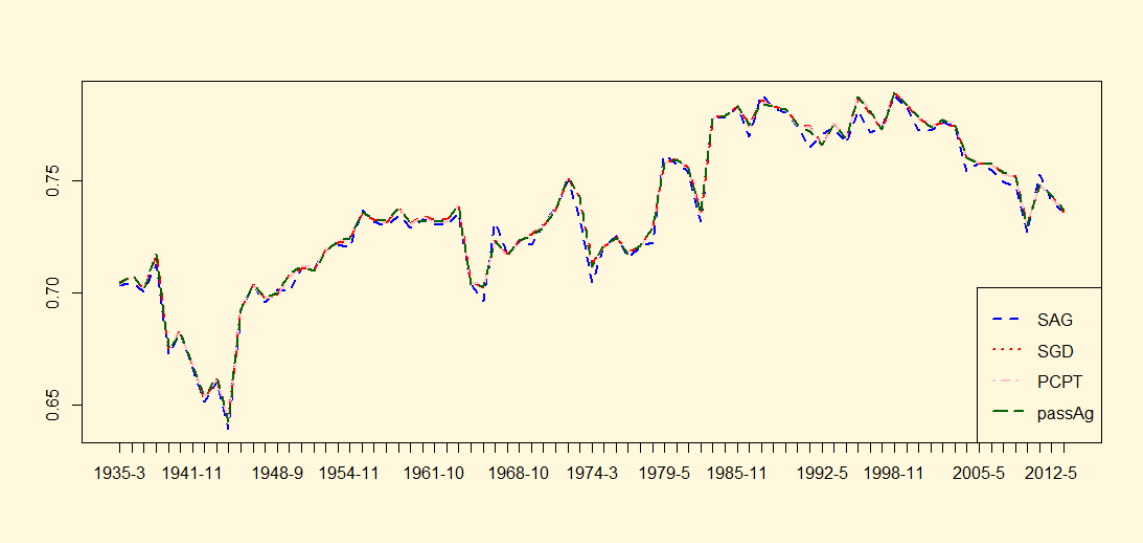

Figure 11: Estimates of parliamentary polarization, by session, by algorithm. Legend abbreviations are logistic regression using stochastic average gradient descent (SAG), stochastic gradient descent classifier (SGD), perceptron (PCPT), passive aggressive (passAg). Notice that performance is essentially identical across algorithms.

Online Appendix E Applying Ensemble Methods

While our primary aim is not to achieve the maximum possible accuracy, one could be concerned that a method with low accuracy was performing unevenly in different time periods for technical reasons unrelated to parliamentary polarization. One way to investigate this is to explore ensemble methods which while more computationally intensive and more difficult to interpret, may achieve higher accuracy. If a more accurate classifier does differentially better in certain time periods—i.e. there are uneven increases in accuracy—it would suggest that our measure of polarization is highly dependent on specifics of the algorithm(s) and thus potentially unreliable. To investigate this, after running the four algorithms mentioned in the paper, we applied gradient boosted trees developed by Friedman (2001) along with an additional regularization parameter as implemented using `XGBoost` (Chen and Guestrin, 2016).

The boosted tree model integrates multiple regression trees in an ensemble. The model is trained additively by starting with one tree and then developing the next tree in such a way as to optimize the objective function (given the residuals from the initial tree), which, as with most machine learning algorithms incorporates both loss and a regularization parameter that penalizes model complexity. For our data we adopt the 'exact greedy algorithm', which first sorts the features according to their importance and then identifies the optimal point at which to make a split for each of these features.

We allow a maximum depth of 14 and use 400 estimators (this was based on a grid search of

these parameters on a previous, similar task), and otherwise adopt the default values, with logistic regression for binary classification as the objective, and learning rate of 0.1. The results are similar to the best of the four algorithms adopted in the paper but shifted up to higher accuracy. The correlation between the two measures is .89. The greatest difference between the two is that the `XGBoost` classifier estimates the WWII years to be even more starkly less polarized than the four algorithms in the paper, and also finds a slightly greater decrease in polarization in the last two decades. Overall, however, the results are very similar and suggest that the overall trend in polarization is stable and not likely to be an artifact of an ineffective classification algorithm. The fit time ranges from 17 to 89 seconds per session when run on 12 cores.

This is particularly reassuring because the `XGBoost` model allows for interactive effects of up to 14 variables (subject to the regularization penalty), which should reassure readers that the results do not strongly depend on words being misinterpreted based on their context or the fact that n-grams were not included in the vocabulary.

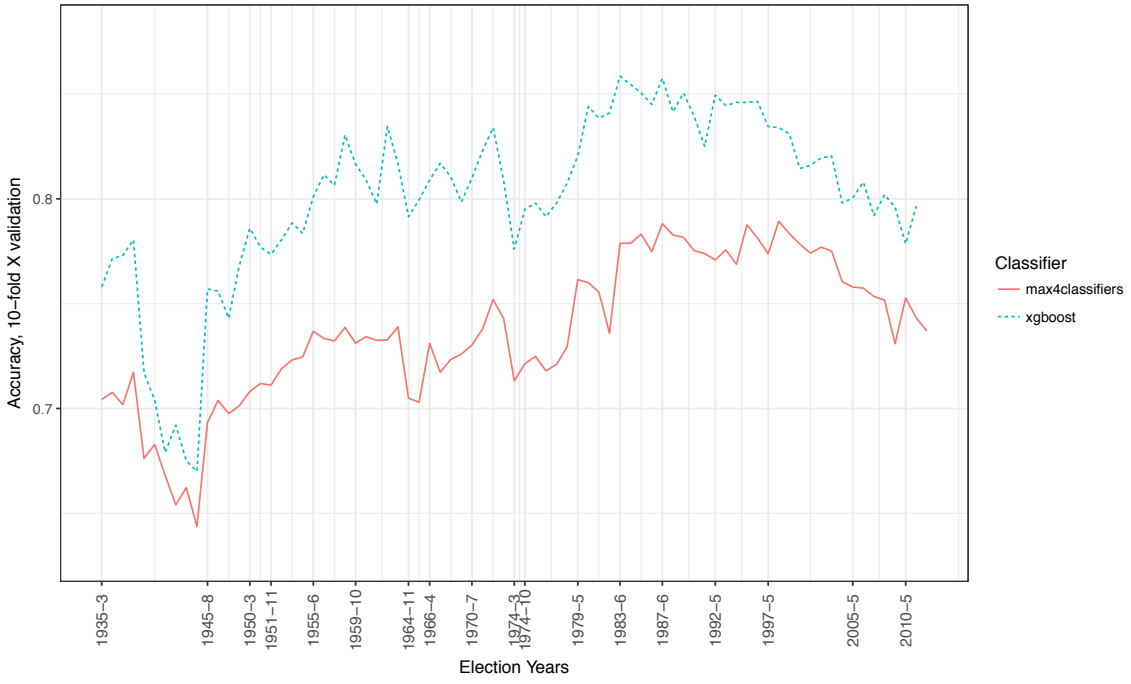

Figure 12: Comparison of Accuracy; Max of Four classifiers versus `XGBoost`

Online Appendix F Further Simulations

To check our variance intuition, we generated 300 members per party, with 10 speeches per person—with each speech generated as discussed in the main body of the text. The variance for individual speakers increases steadily with noise: the mean variance (mean across the 300

different MPs) goes from 0.000003 with no noise, to 0.000008 with 50% noise, to 0.000059 with 90% noise. Of course in an absolute sense there is very little variance in our experiments since the estimates are quite precise with so many speeches, but the principle that individual level variances should grow with noise is correct.

Online Appendix G More on Validation: Roll Calls and other political contexts

Validation of any measurement of a latent characteristic is, of course, non-trivial and we have done our best with the evidence we have. In other contexts, scholars might use other data sources. For example, comparing the output of our approach to the tone of election campaigns (perhaps estimated via models of leaders' speeches on the trail) or the language of newspaper editorials may shed light on its merits. For the US specifically, one might compare our polarization measure with more traditional roll call approaches (in the sense of Barber and McCarty, 2015).

In Westminster systems, as we have explained, validation from legislative voting records is much harder. Nonetheless, in the UK context we do have some work that helps us here. For example, Spirling and Quinn (2010) consider a clustering approach to roll call votes in the UK for the period 1997–2001, and among other findings, they uncover three (latent) groups of MPs: 'Core Loyalists', 'Hardcore Rebels' and 'Mavericks'. In each case they list some particular individuals likely to be part of those sets.

Obviously there are limitations to any comparison: our approach is supervised (rather than unsupervised) and deals in scaling (rather than clustering). Furthermore, our work is predicated on estimating the relative distinctiveness of two parties, rather than factions within one party. Still, we take comfort in noting that we do not draw wildly different conclusions from our findings relative to earlier efforts. To see this, consider Figure 13. There, we have plotted the range of individual positions for a set of legislatures noted by Spirling and Quinn (2010) as being members of the groups they describe. To clarify, we obtain the estimate of the position of a given speech by plugging its characteristics into the function implied by the relevant algorithm. This then gives us a prediction—in terms of that speech's probability of having been made by a Conservative member (recall that the speeches are labelled by the party of the MP making them). Doing this for every speech gives us a set of probabilities for every MP, and we take the mean of a given MP's speech estimates to arrive at a point prediction for the member in question. The top horizontal line in the plot represents the most to the least 'Labour-ish' of the core loyalists (PM Tony Blair, Chancellor Gordon Brown and Home Secretary Jack Straw) mentioned by Spirling and Quinn (2010). Below them, we see the 'Rebels'— including Diane Abbott, Tony Benn, Jeremy Corbyn, Bernie Grant. Notice that they are distinctly 'different' to the loyalists: this makes sense, given they fundamentally disagreed over aspects of policy and direction in government. The bottom

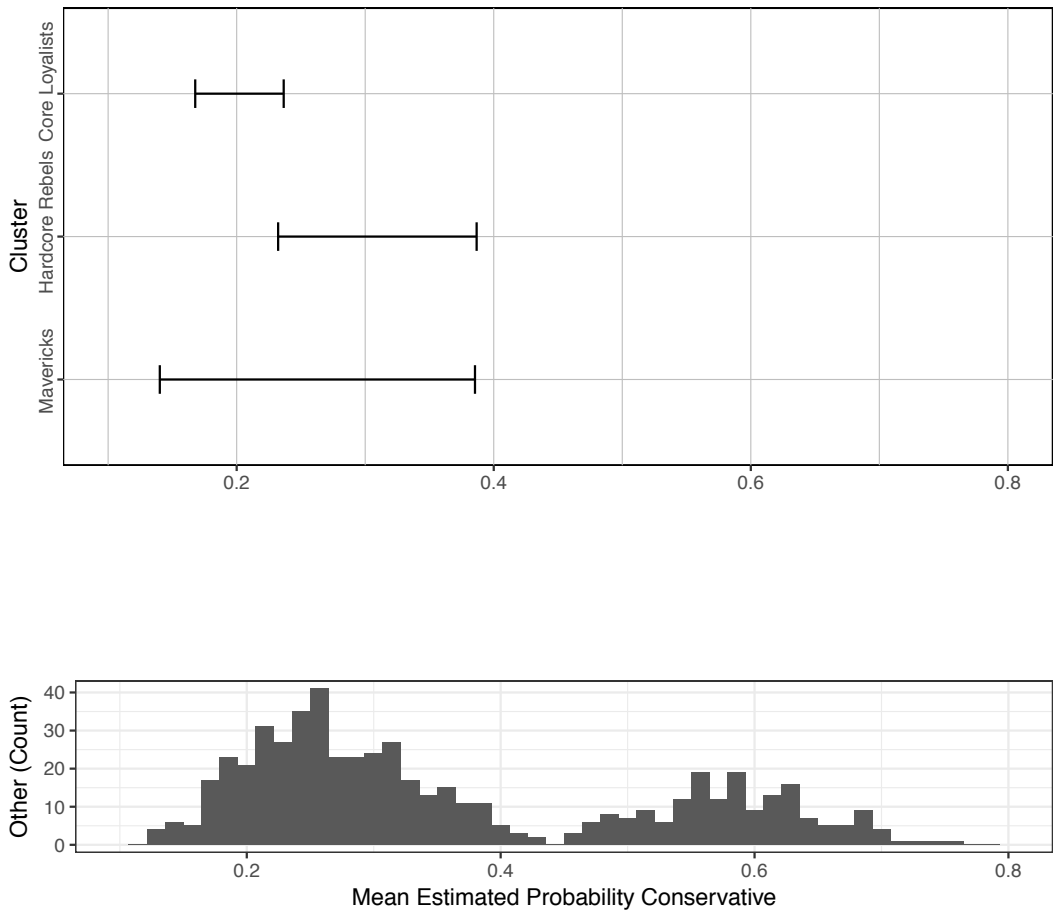

Figure 13: Comparing our individual level scores to various clusters of MPs ('Core Loyalists', 'Hardcore Rebels', 'Mavericks') from Spirling and Quinn (2010). Histogram is of all MPs in 1998.

line represents the 'Mavericks'—such as Tony Banks, Kate Hoey and Denzil Davies—who are hard to pin down in political space: sometimes agreeing with their party bosses, but at other times going out on a limb on policy matters. They have a large spread, exactly as we would expect: sometimes more loyal than the loyalists, sometimes as rebellious as the rebels.

Finally, in the second panel, we provide a histogram for every MP during 1998. We see the two largest blocs, corresponding to the Labour and Conservative parties. Note from the figure that some evidence of the pattern described by Spirling and McLean (2007) for this period—whereby government (left wing) rebels show up not 'left' of the loyalists, but 'right' of them and between loyalists and opposition members—is apparent also in this context. This does not affect the validity of the *aggregate* differences in the sense that there are generally few rebels and they make commensurately small numbers of speeches (and have little to no power over policy), which are not enough to drive the historical patterns. Still, it does suggest some further thought is required before interpreting individual estimates as part of a continuum.

Online Appendix H Model and Data Advice for Practitioners

For our paper, we selected machine-learning algorithms that we suspected would perform well for the data at hand. For users seeking to replicate our style of approach for their own problems, the following practical advice on techniques and data may prove helpful.

To reiterate, we required models that "balance strong predictive power against other concerns such as simplicity, reproducibility, overfitting, and computational time". We would stand by that advice. Ultimately, of course, all of the approaches we used performed similarly. This is not unsurprising given the sheer amount of training data we had; if other users find themselves in similar situations, we would encourage them to choose something that is fast and scalable since the particular technique chosen is unlikely to result in radically different substantive conclusions. When there is sufficient data, it is helpful to also fit a more flexible model that weakens the linearity and monotonicity assumptions to see if such a model still generates similar results, as we discuss in Online Appendix E.

In the event that users of the technique are not so fortunate—that is, they have little training data—we would point them to 'textbook' advice (e.g. Manning, Raghavan and Schütze, 2008) suggesting a general preference for 'high bias' models like Naive Bayes. In addition, for a novel problem, an algorithm that requires little (non-automatic) tuning is probably preferred: so a SVM may be non-optimal at least for an initial run. In the end, of course, we would encourage the use of *several* classifiers. If they produce similar results (specifically in terms of *relative* accuracy over time) which have at least minimal validity, users can be reassured they are probably estimating something useful. If users have weaker priors about

what they expect to find, it may be advisable to steer away from 'black box' techniques that produce results that are generally difficult to interpret: for example, neural networks may not be preferred since while they might produce valid estimates it would be more challenging to identify problems without additional effort.

In terms of data, there are at least three preferred features: first, relatively balanced classes. In our case, we re-weighted to ensure we could compare Conservative and Labour MPs properly, but this had a fairly small effect on our estimates because the share of speeches (and their properties) was similar between the parties over time. Second, a stable vocabulary is preferred. In our case we fix the set of vocabulary based on an initial pass over all the data, but this would not work if there was not substantial overlap in the words used on the documents.[1] In any case, we suggest users examine possible changes to vocabulary size in the sense suggested by Gentzkow, Shapiro and Taddy (2016). Third, we require relatively consistent amounts of noise. That is, to the extent that term-use predicts partisan affiliation, the strength of that signal should be as constant as possible over time. If it is not, there is a danger that claims that a system has shifted to a period of low polarization are based on members simply saying more non-partisan 'filler' words, even if the underlying division over substantive terms has not changed (or indeed, has become more stark). That said, our simulations suggest that noise begins to affect the polarization measure only at high levels (e.g. greater than 90%), although this naturally also depends on the distinctiveness of the vocabulary of the parties and the amount of training data available (Figures 1, 2).

One final suggestion for evaluating the performance of the approach comes from Niels Goet. He suggests studying the autocorrelation between the accuracy estimates for the time periods. Since the model is fit separately on each session, these could in theory differ radically. But if the method consistently identifies the notion of polarization, then the autocorrelation should be large. This is because we know from the literature that political polarization in legislatures does not change overnight from say, very high to very low. Thus, if the correlation between sessions is very small, erratic, or the measure is constant, we likely have a failure of the approach to reliably detect the signal.

[1] This is unlikely but could happen if the text data was on radically different topics.

References

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.

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.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* pp. 1189–1232.

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

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*

Jensen, Jacob, Suresh Naidu, Ethan Kaplan, Laurence Wilse-Samson, David Gergen, Michael Zuckerman and Arthur Spirling. 2012. "Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech [with comments and discussion]." *Brookings Papers on Economic Activity* pp. 1–81.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

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

Spirling, Arthur and Kevin Quinn. 2010. *Journal of the American Statistical Association* 105(490):447–457.