

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

Andrew Peterson\*

Arthur Spirling<sup>†</sup>

---

\*Postdoctoral Researcher, University of Geneva. [andrew.peterson@unige.ch](mailto:andrew.peterson@unige.ch)

<sup>†</sup>Associate Professor of Politics and Data Science, New York University. [arthur.spirling@nyu.edu](mailto: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 cyclicalality 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.

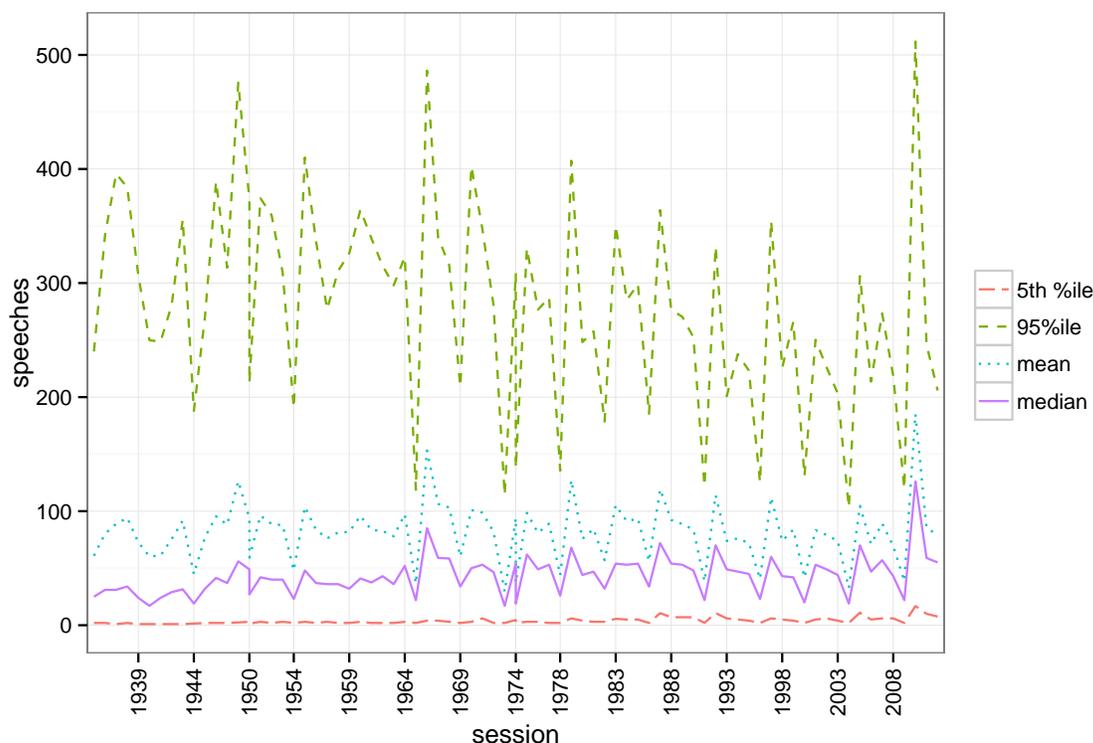


Figure 6: Number of Speeches By Member Per Session.

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.

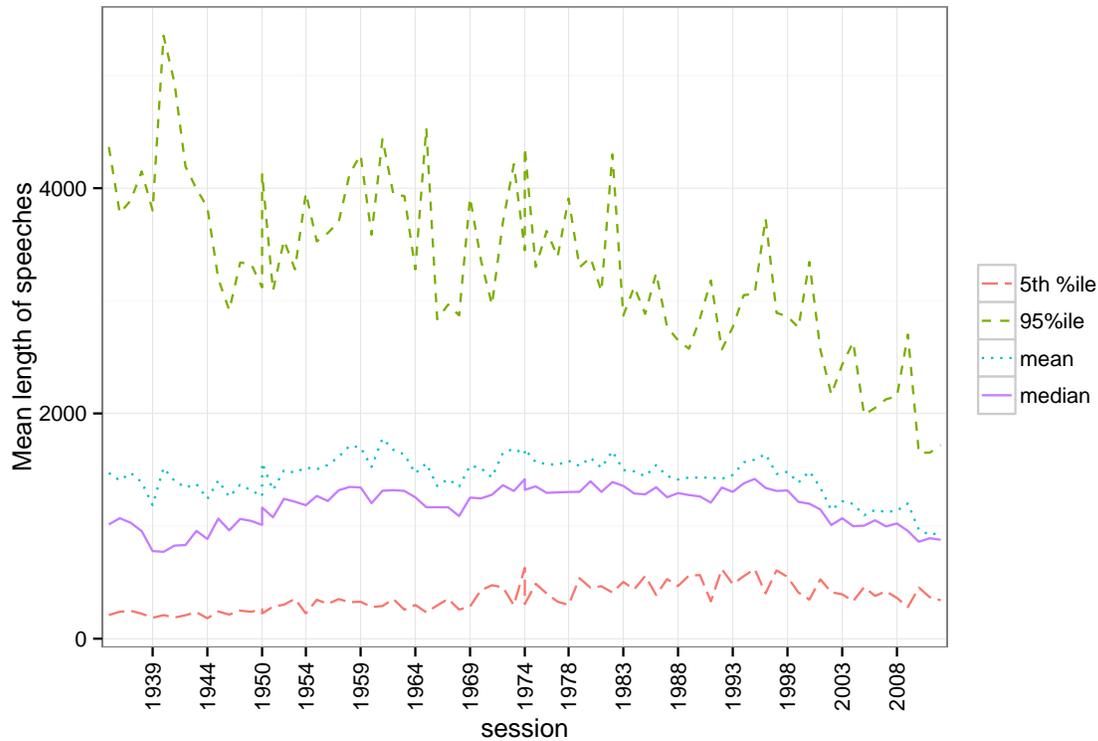


Figure 7: Mean Length of Speeches By Member Per Session.

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.

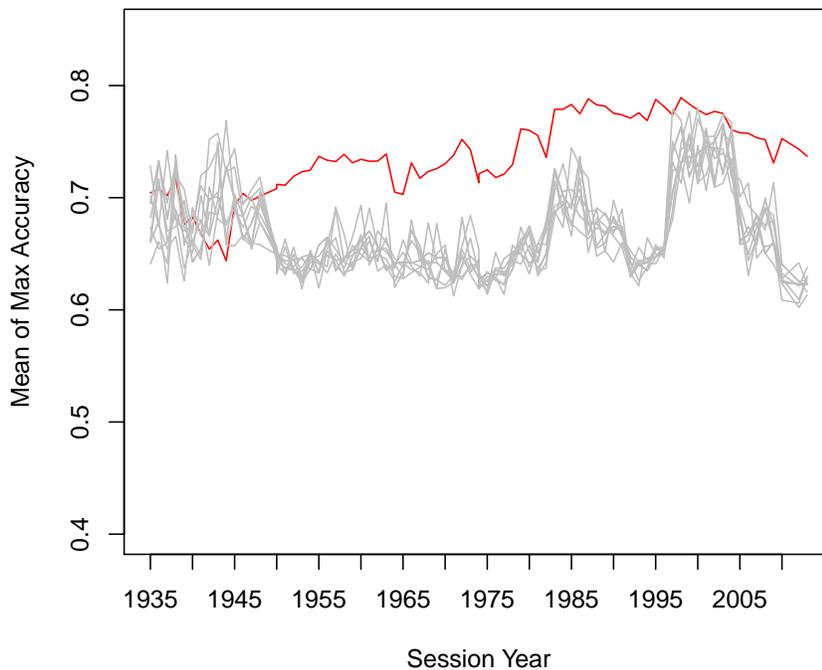


Figure 8: Estimates of parliamentary polarization, by session, by algorithm. The accuracy using real party labels (our polarization measure) is in red, while 10 runs with party labels randomized by speaker are presented in grey.

## 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.

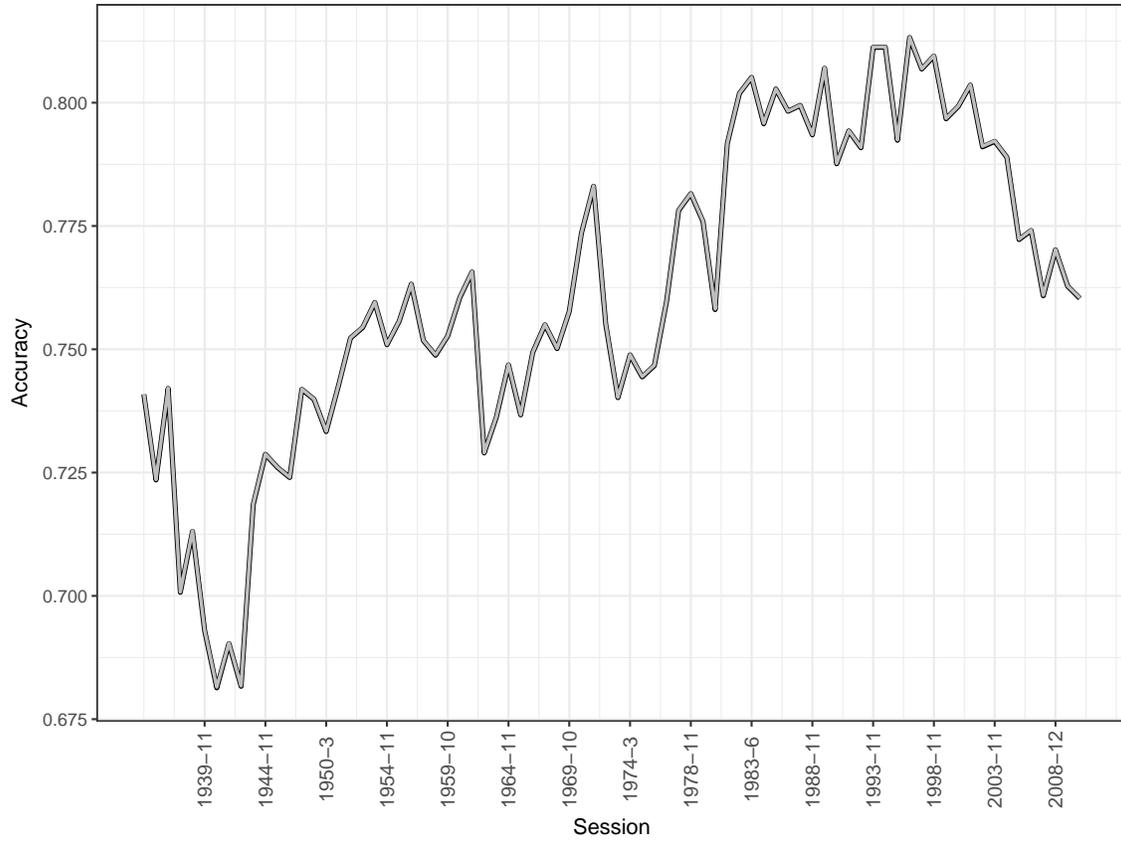


Figure 9: Polarization measure with bootstrap confidence interval, based on resampling texts within each session

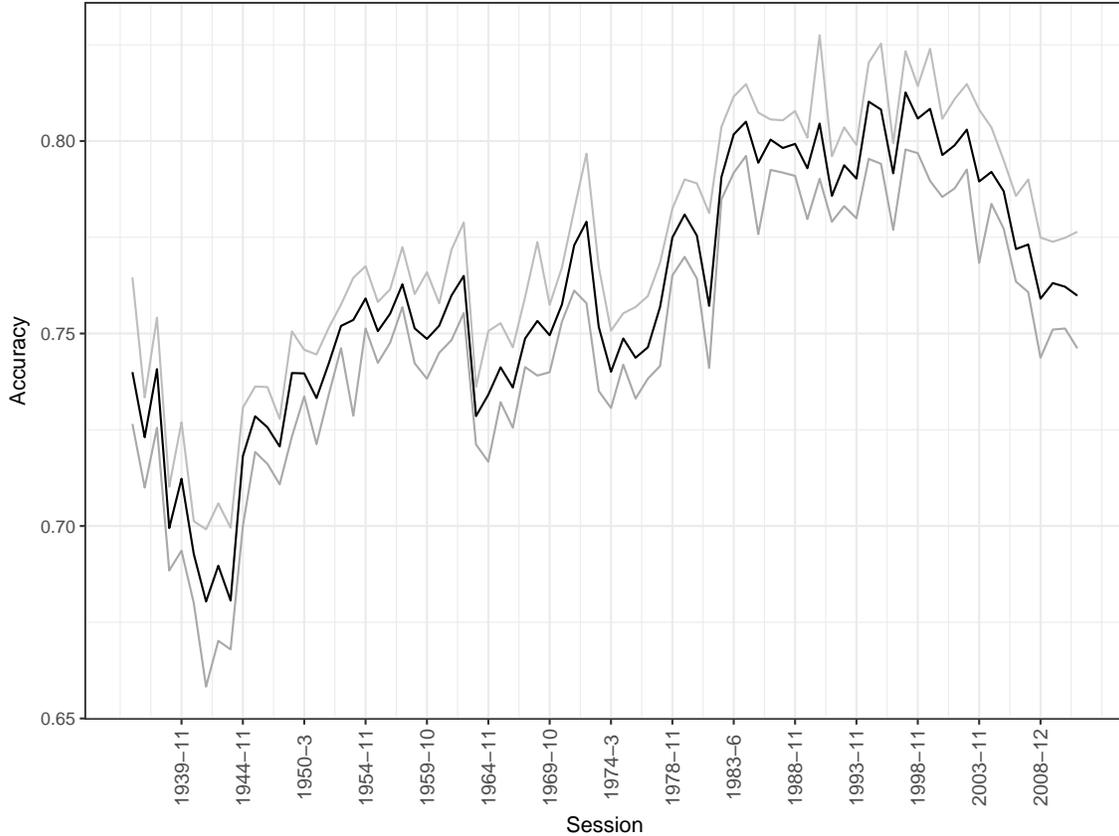


Figure 10: Polarization measure with percentile bootstrap confidence interval, based on all estimates from each of the four algorithms, while resampling texts within each session

## 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.

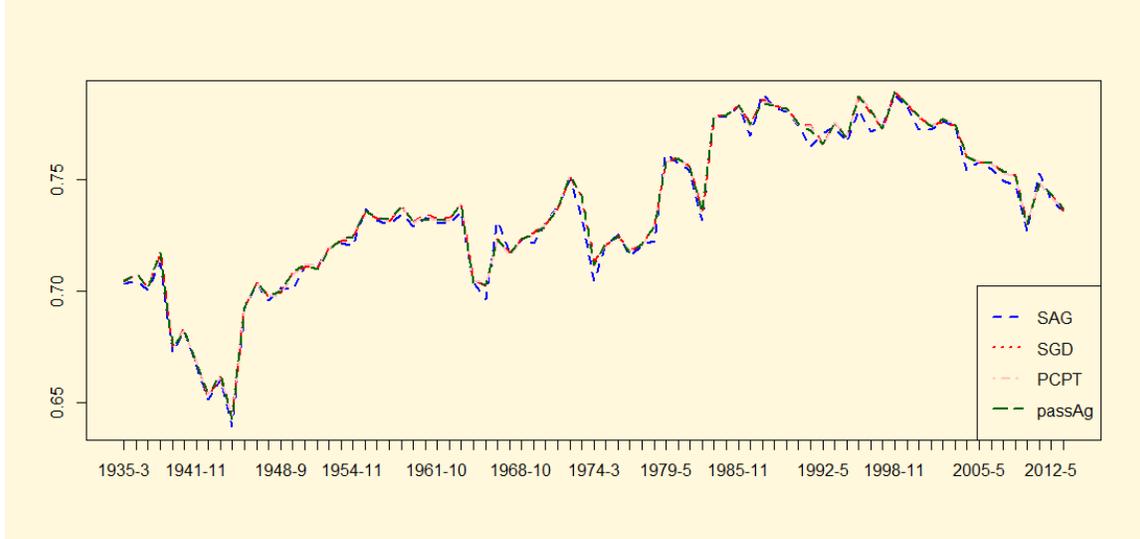


Figure 11: Estimates of parliamentary polarization, by session, by algorithm. Legend abbreviations are logistic regression using stochastic average gradient descent (SAG), stochastic gradient descent classifier (SGD), perceptron (PCPT), passive aggressive (passAg). Notice that performance is essentially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

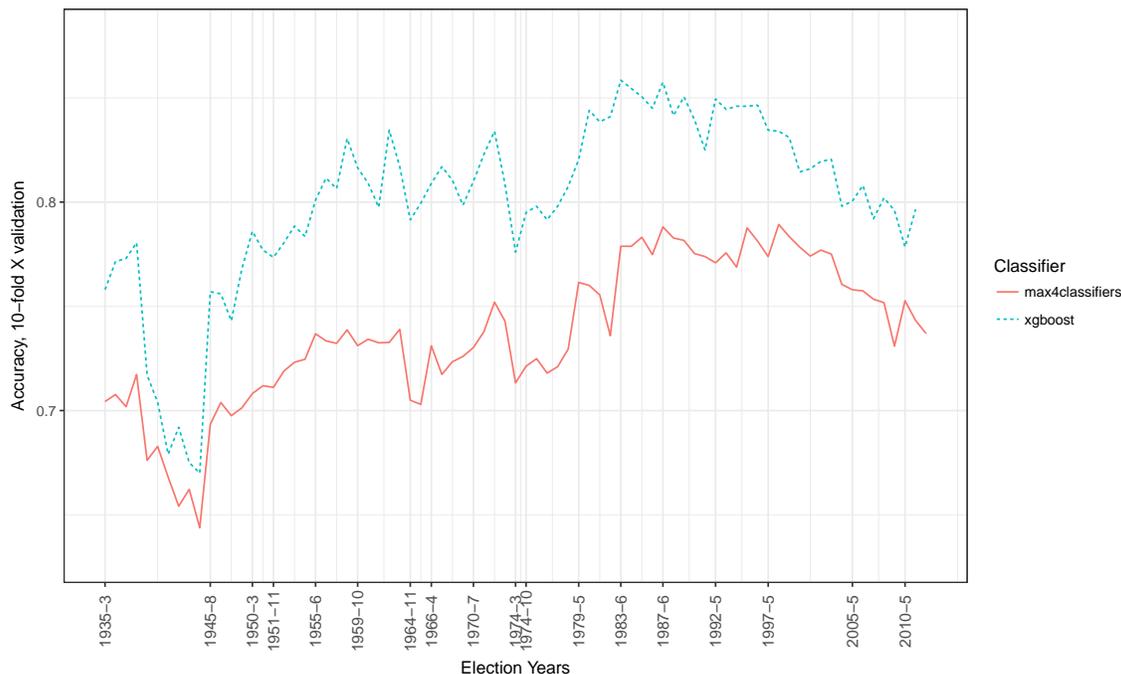


Figure 12: Comparison of Accuracy; Max of Four classifiers versus XGBoost

## 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

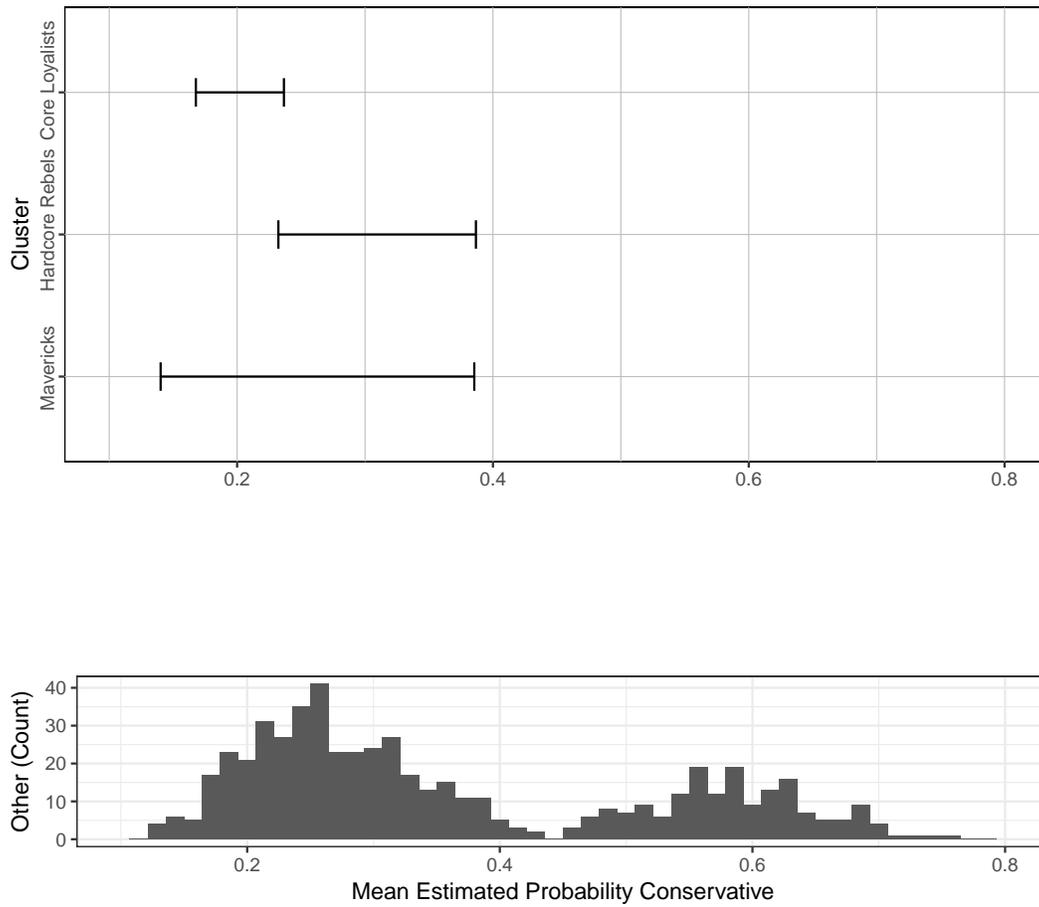


Figure 13: Comparing our individual level scores to various clusters of MPs (‘Core Loyalists’, ‘Hardcore Rebels’, ‘Mavericks’) from Spirling and Quinn (2010). Histogram is of all MPs in 1998.

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.<sup>1</sup> 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.

---

<sup>1</sup>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.