Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015

Niels D. Goet

Data Scientist, Inspera AS, Oslo, Norway. Email: ndgoet@gmail.com

Abstract

Political scientists can rely on a long tradition of applying unsupervised measurement models to estimate ideology and preferences from texts. However, in practice the hope that the dominant source of variation in their data is the quantity of interest is often not realized. In this paper, I argue that in the messy world of speeches we have to rely on supervised approaches that include information on party affiliation in order to produce *meaningful* estimates of polarization. To substantiate this argument, I introduce a validation framework that may be used to comparatively assess supervised and unsupervised methods, and estimate polarization on the basis of 6.2 million records of parliamentary speeches from the UK House of Commons over the period 1811–2015. Beyond introducing several important adjustments to existing estimation approaches, the paper's methodological contribution therefore consists of outlining the challenges of applying unsupervised estimation techniques to speech data, and arguing in detail why we should instead rely on supervised methods to measure polarization.

Keywords: parliamentary debate, UK House of Commons, polarization, text analysis

1 Introduction

Parliamentary debate is, at its core, an expression of political conflict. Members of parliament (MPs) use the plenary to express their concerns, debate proposals, and communicate their stance on issues. The parliamentary debating arena, with its host of diverse expressed views, is a promising venue for measuring the political stance of MPs. This is especially the case since the (still) predominant roll call-based approaches do not travel well in parliamentary systems such as the UK (Spirling and McLean 2007; Hix and Noury 2010). Here, voting is driven by career incentives and government allegiance rather than by preferences on specific bills (Benedetto and Hix 2007; Kam 2009). Voting against the party, then, must be seen as the "nuclear option" (Proksch and Slapin 2015). Conversely, the rules of debate of the UK House of Commons—contained in the Standing Orders—grant MPs significant freedom to participate in debates.[1] As long as MPs vote with the party line, legislative speeches remain relatively unconstrained (Schwarz, Traber, and Benoit 2017).

Author's note: I would like to thank Andrew Peterson, Max Goplerud, David Doyle, Radoslaw Zubek, Simon Hug, and Kamil Marcinkiewicz for helpful suggestions and comments on earlier drafts. I am also grateful to discussants and audiences at the American Political Science Association meeting (2017) and the ECPR Standing Group on Parliaments (2017), who provided helpful feedback. My manuscript benefited greatly from detailed comments from three anonymous referees, and from the editor at *Political Analysis*. The usual disclaimer applies. Replication materials are available on the *Political Analysis* Dataverse (Goet 2018). Supplementary materials for this article are available on the *Political Analysis* website.

[1] Although note that we cannot expect speeches to be entirely free of "selection effects". On the one hand, this may be a function of the party leadership that wishes to protect the party brand (cf. Proksch and Slapin 2015). On the other, research has shown legislators in the UK may self-select into speaking. Herzog and Benoit (2015) for example find that MPs from constituencies that face economic hardship are more likely to use partisan speech to challenge austerity. And in the US context, Grimmer (2013) shows that legislators with extreme preferences are more likely to speak in policy debates while moderates stick to credit claiming, artificially boosting levels of polarization. I aim to investigate the effect of such dynamics on the measurement of polarization in future work.

***Political Analysis* (2019)**
vol. 27:518–539
DOI: 10.1017/pan.2019.2

Published
29 April 2019

Corresponding author
Niels D. Goet

Edited by
R. Michael Alvarez

In this paper, I propose that we should therefore focus on the aggregate level and use speech data to consider polarization[2] between political parties. Political scientists have had a long tradition of applying unsupervised measurement models to gauge ideology and preferences from texts. However, the hope that the dominant source of variation in their data is related to the phenomenon they want to measure is often not realized. My main argument is that we should therefore rely on *supervised* estimation methods that include information on party affiliation. In contrast to their unsupervised siblings, such supervised models attempt to identify which speakers use a vocabulary that is similar to speakers from one versus another party, ensuring that variation in word use is related to a stable construct. To support this argument empirically, I apply supervised and unsupervised models to 6.2 million records of parliamentary speeches from the UK House of Commons (1811–2015).

The paper makes three contributions to the "text-as-data" literature. First, and most importantly, it addresses a long-standing debate in political science about the usefulness of supervised versus unsupervised approaches in text analysis with application to political preferences. Second, this paper presents a coherent framework for evaluating the performance of text-based measures of polarization, building on important work that seeks to challenge the often-applied strategy of conducting "some form" of validation (e.g. Quinn *et al.* 2010). Third, the paper makes a comparative assessment between one of the most commonly used text scaling models—`Wordfish` (Slapin and Proksch 2008)—and a novel machine classification approach (Peterson and Spirling 2018).[3] Through an application of both approaches over an extended time frame, this paper provides a comprehensive overview of the choices and pitfalls that researchers face when they apply text analysis tools to measure polarization from parliamentary speeches.

The relevance of the work presented in this paper extends beyond the text analysis field. Many theories of institutional design incorporate some element of polarization (e.g. Binder 1996; Diermeier and Vlaicu 2011). The validation framework and applications presented here, as well as their evaluation, should enable researchers to select an appropriate measure of polarization to test a range of hypotheses from this literature in the UK context, as part of the new research agenda of British Political Development (BPD) (Spirling 2014).

2 Measuring Preferences Using Text

Until recently, the "text-as-data" approach to measuring political preferences focused on a narrow set of texts such as party manifestos (e.g. Laver, Benoit, and Garry 2003; Slapin and Proksch 2008).[4] Recent advances in the processing of text data, the digitalization of records, and the development of new algorithms, have enabled researchers to shift the focus to parliamentary speeches, and the preferences of individual legislators (e.g. Lauderdale and Herzog 2016; Schwarz, Traber, and Benoit 2017). This is an important innovation because, as pointed out above, the conventional roll-call votes-based approach to measuring ideal points does not travel well in the parliamentary context (see also Vandoren 1990; Carrubba *et al.* 2006; Carrubba, Gabel, and Hug 2008; Hug 2010; Schwarz, Traber, and Benoit 2017). Here, the use of *speeches* to infer the ideological standpoint of legislators has two main advantages (cf. Proksch and Slapin 2015, 7). First, speeches are less subject to partisan control than voting. Defection on votes can be seen as the ultimate act of defiance. In contrast, speeches afford MPs the opportunity to express dissent in a way that is

[2] Polarization is defined as the degree to which MPs are ideologically proximate to one another, i.e. how consistently MPs fall within their respective parties across policy issues.

[3] One might note that the two methods in this paper are fundamentally different. `Wordfish` is unsupervised while the machine classification approach applied in this paper is supervised. However, while there may be more comparable methods available (Monroe, Colaresi, and Quinn (2008) comes to mind), `Wordfish` is a popular and highly cited method for analyzing text, and classification accuracy approaches are relatively less common in political science. `Wordfish` is the natural tool that applied researchers would reach for, and therefore is the relevant benchmark for comparison.

[4] For a comprehensive overview of the use of computational text analysis in legislative studies, see Proksch and Slapin (2014).

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

less likely to harm their own or their party's position. Second, even if such partisan control is not exercised, votes reduce an actor's preferences to one of three options—in favor, against, or abstain—whereas speeches enable MPs to express their views in a more nuanced way.

Challenges

Yet, the use of text to measure political preferences certainly does not come without problems. Researchers need to account for the high-dimensional nature of speech data, i.e. the fact that not all phrases used by political actors, or by one such actor in a single speech, map onto the same latent concept. Traditionally, supervised Poisson scaling algorithms such as `Wordfish` have focused on election manifestos. Political actors spend a considerable amount of time and effort editing such documents to ensure that the topic of each part of the text as well as the message that it carries are beyond doubt. Consequently, researchers can easily identify the policy area that each section covers. In contrast, the world of debate and speech is much messier. Speeches are not exclusively related to the bill or item on the agenda. Rather, politicians often go off-topic, speaking to different matters, or combining their statement on the discussion topic with several other messages. Moreover, in most unsupervised scaling applications (e.g. Lauderdale and Herzog 2016; Schwarz, Traber, and Benoit 2017) speeches are aggregated by actor at an appropriate level (e.g. for each debate), which means we potentially include many statements that are unrelated to the policy dimension that we are interested in.

Dimensionality becomes particularly problematic when applying unsupervised scaling models like `Wordfish`, which recover only one of many possible latent dimensions that can be extracted from speech data. We do not know *a priori* if that dimension—i.e. the axis that explains the largest degree of variation in relative word frequencies—maps onto the kind of political conflict that we want to measure. Moreover, language itself undergoes significant alterations over a 200-year time period. For example, the world "welfare" implied something very different in the 1800s than in the 1950s, when the welfare state as we know it today was constructed. As will become clear in the empirical analysis, the supervised and unsupervised approaches deal with this problem to varying degrees of success.

Validation

How then, can we assess whether a supervised or unsupervised model produces better results? Unsupervised methods come with significant post hoc validation costs, as the researcher "must combine experimental, substantive, and statistical evidence to demonstrate that the measures are as conceptually valid as measures from an equivalent supervised model" (Grimmer and Stewart 2013, 5). There are important examples of "best practice" in the validation of text-based measures of political phenomena (see in particular Quinn *et al.* 2010, who present a comprehensive framework for evaluating estimates of political attention (or: "topics")). However, as it stands, there is no consistent framework for evaluating text-based measures of *polarization* with a set of simple, predefined tests that researchers can follow. This is problematic because, as researchers, we want to be able to compare the performance of several different methods against common standards. Such a framework needs to combine rigor with speed and ease: when evaluating a speech-based measure over a long time period (in this case, over 200 years), the post hoc validation needs to give us confidence in the results. At the same time, it should limit the costs of the validation exercise to allow us to reap the benefits (i.e. speed) of using an unsupervised method.

The framework for evaluating text-based measures of polarization developed in this paper is summarized in Table 1. It is divided into three types of validity (face, convergent, and construct), and includes several tests for each type.[5]

[5] For a discussion of different types of measurement validity in political science, see Adcock and Collier (2001).

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Table 1. Validation framework.

Test	Key question	Test
1. Face validity		
1.1 General	Is there a reasonable level of stability of estimates from session to session within parliaments?	Visual
1.2 Detailed	Do the estimates correspond with *a priori* expectations of levels of polarization?	Visual
2. Convergent validity		
2.1 Session level	Do the session-level estimates of polarization correspond with a comparable exogenous measure?	Correlation
2.2 Individual level	Do the individual-level estimates correspond with the self-placement of MPs?	Correlation
3. Construct validity		
3.1 Between-session consistency	Are the individual-level estimates consistent from session to session?	Correlation of MP positions from t to $t+1$
3.2 Individual-level distribution	Does the distribution of individual-level estimates show a reasonable degree of separation by party?	Visual
3.3 Explanatory power of the party label	Does the observed variation in individual-level estimates explain a reasonable proportion of the variation in party labels?	R^2 from regressing individual estimates on party labels

Face Validity: First, we consider face validity, which includes both a general and a detailed benchmark. The general test is a quick and simple impression of the *distribution* of the estimates over time:

> **1.1 General test**: The level of variability between sessions within the same parliament should be at a reasonable level: although we may expect some variation from session to session, we should not observe an at-random pattern of switches between high and low polarization from year to year, and especially within one parliamentary term.

However, as (Quinn *et al.* 2010, 216) rightly note, "[f]ace validity is inherently subjective, generally viewed as self-evident by authors and with practiced skepticism by readers". This does not mean that face validity cannot be useful, if applied in a consistent way. A second, and more important face validity criterion of my framework therefore is that the measure must pass a detailed "historical test":

> **1.2 Detailed test**: Outliers in our estimates should correspond with *a priori* expectations derived from authoritative (secondary) sources.

This test requires the researcher to, *prior to estimating the measure*, set out testable and specific expectations of historical outliers (low or high polarization) on the basis of authoritative secondary sources. In the UK case, we should observe an up-going trend in the levels of conflict over time, and especially after the 1880s as party organization takes hold. Specific periods of low and high polarization are outlined in Table 2.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Table 2. Detailed test (1.2): key outliers in polarization in the UK Parliament.

id	Period	Event	Polarization
1	1815 and 1846	The Corn Laws of 1815 sparked a period of polarization as tensions between government and opposition rose. In 1846, the Corn Laws were repealed under Sir Robert Peel, marking a return to normality and lower levels of conflict (McLean and Bustani 1999).	High
2	1832	Great Reform Act, which sought to extend the vote and disenfranchise "rotten boroughs" (Cox 1987).	High
3	Around 1880	Irish obstruction dominates the parliament's agenda, sparking fierce conflict between political parties (Pugh 1982).	High
4	1885–1886	Conflict over the first Irish Home Rule Bill that PM Gladstone tried to pass, only to resign when Parliament failed to do so (Pugh 1982).	High
5	1906	Liberal welfare reforms under David Lloyd George causes increased polarization between government and opposition MPs (Pugh 1999).	High
6	1914–1918	National coalition: WWI (Pugh 1999).	Low
7	1923	Coalition government: MacDonald's Labour administration that governed with support from Asquith's Liberal party (Pugh 1999).	Low
8	1939–1945	National coalition: WWII (Pugh 1999).	Low
9	1979–1990	Tensions between the opposition and government parties over Thatcher's liberalization agenda (Pugh 1999).	High
10	2010–2015	Conservative–Liberal Democrat coalition.	Low

Convergent Validity: Second, the framework considers how well our estimates converge with results obtained with supervised methods, at two levels. First, at the aggregate level (here defined as the yearly parliamentary session), we should expect that levels of polarization correlate with other exogenously defined measures:

> **2.1 Session-level test**: The level of polarization in sessions should correspond well with exogenously defined measures of polarization.

The second level considers the MP-level estimates:

> **2.2 Individual-level test**: The positions of MPs should correlate with their own left–right placement from an exogenous dataset.

As speeches are a direct reflection of the political preferences of legislators (and to some degree of political constraints), we should observe a relatively high correlation between self-placement and our estimates.

Construct Validity: Third, the framework evaluates three measures of consistency relating to construct validity. First, we evaluate the variation in the position of MPs from session to session. We may expect the ideal point of legislators on a one-dimensional scale to vary somewhat because the agenda will include different items for each session. However, legislators should otherwise remain relatively consistent in their overall position across the issues discussed over the course of a parliamentary session:

> **3.1 Between-session consistency**: The positions of MPs should correlate at a reasonable level between successive sessions.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Second, a visual test of individual-level scores can give some indication of the performance of a measure:

> **3.2 Individual-level distribution**: The Empirical Cumulative Distribution Function (ECDF) of individual-level estimates should show a reasonable separation of parties *and* key individuals should be placed as expected.

In this context, "reasonable" primarily refers to: (i) a separation where individuals from the left-wing party are not found to the right of the right-most member of the right-wing party on the political spectrum; (ii) individuals do not drastically change position from one session to the next.

Third, the individual-level estimates should account for a reasonable proportion of the variation in party labels:

> **3.3 Explanatory power of the party label**: The variation in individual-level estimates should be a good predictor of the party label for each session.

We can assess this third test by regressing individual estimates on the party label and taking the R^2 from the model. We want this proportion to reflect some degree of party control: i.e., the individual estimates should not be assigned at random. At the same time, some unexplained variation should remain as, for reasons elucidated above, the party should not structure speech completely.

Applying the evaluation scheme

The evaluation scheme has two applications—both of which will be used in this paper. First, it may be used to evaluate, comparatively, the merits of two or several different approaches to measuring polarization based on text. Second, it can be applied to establish whether or not a measure is "valid enough" to be used in other, more substantive applications.

When can we safely assume that a measure that we have generated with our text analysis algorithm has produced a "valid" measure? This is largely dependent on the time frame. For the application in this paper, the "threshold" is as follows:

> **Threshold**: The measure should pass *all* tests of the evaluation scheme, but not for all sessions included in the analysis.

For a measure that spans 200 years, the general paucity of comparable measures of polarization means that some of the tests can only be conducted across a smaller time scale. The (obviously wrong but practically useful) assumption made here is that, if the test performs well for any session it should also perform satisfactorily for any other randomly sampled session. A "pass" then, is defined as satisfying a minimum level *and* performing better *in comparison with other measures*.

3 Case Selection & Data[6]

The UK provides a promising institutional setting to develop text-based measures of polarization. Its legislative process affords MPs ample opportunity to voice policy-related opinions—the kind of statements that we expect to reveal ideological preferences. Today, the legislative process for public bills in the Commons consists of three readings, and includes six distinct stages. After presentation (stage 1: first reading), each bill is subject to a general debate (stage 2: second reading), and after a committee stage (stage 3), a detailed debate ("report stage", stage 4) during which MPs discuss the committee's amended bill and may propose amendments, followed by a final third reading (stage 5) at which the final version of the bill may again be debated (but no

[6] Replication materials are available on-line as Goet (2018).

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

amendments may be proposed) (see Standing Orders of the House of Commons—Public Business 2016, arts. 57–83).

At the end of the third reading, a vote is taken whether to approve the bill. This stage is followed by a similar set of readings in the House of Lords. After the third reading in the Lords, the bill is sent back to the Commons, giving the latter the opportunity to debate and review the Lords' amendments, and to propose their own (stage 6). After this stage, the bill may receive "royal assent", bringing it into effect. In sum, there are no fewer than *four* opportunities for legislators to engage in plenary debate on a bill (stages 2, 4, 5, and 6).

Over the entire period studied in this paper, members of parliament remain relatively free to engage in debate, thereby avoiding the problems of nonvoting and selection that undermines roll-call based analyses. Most changes to curb the speaking rights of MPs were, over the time period studied, introduced at the macrolevel of agenda rights or the timetable.[7] For an MP to speak, it suffices for her to rise from her chair, after which the Speaker may give her the floor. Thus, it is the Speaker—by all intents and purposes a neutral institution—who decides who speaks; not the party. Moreover, members may submit amendments freely at the report stage, giving them ample opportunity to put forward their views.

To implement the text algorithms that I shall outline below, I use newly collected data from the UK House of Commons *Hansard* archives. The dataset includes 6,224,352 speeches from 1811 and 2015, spanning 233 parliamentary sessions from the 5th session of the 4th Parliament, up until and including the final session of the 55th Parliament.[8] Details of the data gathering process, preprocessing decisions, and a procedure for removing procedural phrases are provided in Appendix A in the on-line supplementary materials.

4 Unsupervised versus Supervised Models

The core difference between unsupervised and supervised models—and the reason why we may expect them to produce different results—lies in the way they use variation in data to yield estimates. Unsupervised models attempt to describe variation in word use. The scores that these models produce identify which speakers tend to use similar words to one another, whether overall or within debates. These estimates may or may not prove to have anything to do with the party, or indeed to have any stable structure over different debates or sessions. The results could be completely confounded by different speakers talking about different topics, or other sources of variation in word use. By contrast, the supervised variant is designed to find variation in word use that predicts party labels. The scores from the supervised models identify which speakers tend to use similar words to speakers from one party versus speakers from the other party. These estimates are guaranteed to be driven by the "party factor", regardless of the number of topics that are addressed, or other sources of variation in word use. When dealing with the messy world of text data, supervised approaches should therefore be expected to perform better than their unsupervised siblings.

Unsupervised models: Poisson scaling

The first two models applied in this paper are of the unsupervised variant, and are adapted forms of the `Wordfish` algorithm. This Poisson scaling model is the most appropriate for our purposes.[9] First, recent applications have successfully employed the model to the study of ideal

[7] These include, for example, the introduction of a lottery system for tabling bills (early 1800s), the creation of the closure rule (1882) and its reform (1887), as well as Balfour's railway timetable reforms (1902), intended to streamline the agenda and limit speaking time. Even the time of individual members' speeches is largely beyond party control, as the only formal limit was introduced in 1988, and relates to an explanatory speech upon a motion to move the adjournment for the purpose of discussing a matter of urgency (see the amendment of Standing Order 10(1) of 27th February 1986).

[8] This figure includes all speech acts, i.e. questions, interpellations, and speeches.

[9] For a validation of the model's assumptions in the context of the data used in this paper, see Appendix B in the on-line supplementary materials.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

points in legislatures, when applied at the *debate level*. Schwarz, Traber, and Benoit (2017), for example find in a study of the Swiss legislature based on an energy policy debate (2002–2003) that speech-based estimates reveal larger differences of ideology within parties than roll call-based measures. In turn, Lowe and Benoit (2013) find a high correspondence between human coding of texts from the austerity budget debate in the Irish legislature (2009) and `Wordfish` estimates. Second, although some have argued that correspondence analysis (CA) (Greenacre 2016)—the least-squares sibling of `Wordfish`—is a more appropriate technique (Lowe 2013), in practice, `Wordfish` is more robust to outliers in word use (Lauderdale and Herzog 2016).[10]

`Wordfish` is a statistical model that allows users to estimate a latent position of an actor on the basis of word frequencies (Slapin and Proksch 2008). Its introduction marked an important advantage over previous techniques, as it allows for time series estimates, does not require reference texts, and relies on the words used in each document and provides the contribution of each term to a given estimated position of that document. Each text included in the `Wordfish` model is treated as a separate position, and all positions are estimated simultaneously. Crucially, as `Wordfish` is an unsupervised approach, we do not include prior information about the position of the actors in the model. When applied to collections of speeches by individuals in a legislative session—as opposed to the original application to party manifestos over time—we therefore obtain a distribution of latent positions of those legislators *vis-à-vis* one another in a one-dimensional space.

Dimension-level scaling

`Wordfish` lives by the assumption of unidimensionality: it assumes that the principal dimension extracted from the texts represents their political content. If we wish to know the position of actors on a specific policy dimension such as the economy, we would therefore have to run the model on texts where we know *ex ante* that they express the actors' views about the economy (Slapin and Proksch 2008, 711). Indeed, the main challenge when estimating ideological positions from speeches with `Wordfish` is to pin down the dimension of interest. *Ex ante*, we need to minimize variation in word use that relates to dimensions other than the one we wish to extract. In the model's original application to manifestos this proved relatively straightforward as titles and subtitles in written manifestos allow for classification of topics; with speeches, political actors are more likely to go off-topic and dedicate (parts of) their speech to matters not related to the (policy) dimension that is on the agenda. Simply looking at the titles as recorded in *Hansard* will therefore not yield the desired result. The first strategy applied in this paper deals with this dimensionality problem by including a dictionary, classification, and scaling stage.

Prior to the estimation stage, I create a dictionary (stage 1) and apply a classification algorithm (stage 2) to identify speeches that are related to one specific policy area (or dimension): the *economy*. At the first stage, I construct a dictionary of economy-related terms on the basis of the Comparative Manifesto Project (CMP, Volkens *et al.* (2016)) dataset, complemented with a number of terms specific to the UK context. Subsequently, at the second stage, I train a stochastic gradient descent (SGD) classifier to these data, and use the trained model to sort all speeches into the economic and noneconomic categories.[11] At the last and final stage, I apply `Wordfish` to recover

[10] The two estimation strategies outlined below are implemented with a dataset of speeches where: (i) procedural phrases have been removed; (ii) only discussions with five speakers or longer are included; (iii) speeches are fifty words long at a minimum; and (iv) debates only include speeches by those parties that account for at least one fifth of speeches made in the discussion. The latter choice is made to limit the drop-out rate of terms when reducing the sparsity of the matrix.

[11] Appendix D in the on-line supplementary materials contains details of the dictionary, and of the selection and classification procedures.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

estimates for the position of legislators on a specific policy dimension within a parliamentary session.[12]

Debate-level scaling: `Wordshoal`

A second approach to deal with the high-dimensional nature of text data is to estimate legislator positions at a more granular level (i.e. within debates) and devise an appropriate way to aggregate estimates across different axes of conflict. This is the solution offered by Lauderdale and Herzog's "`Wordshoal`" (2016). Here, we first use the standard unidimensional Poisson scaling model `Wordfish` (Slapin and Proksch 2008) to estimate debate-specific legislator positions, and subsequently apply Bayesian factor analysis on the sparse matrix of debate-specific legislator positions on each debate to recover their latent position.[13]

At the first stage of the estimation, we have to establish what constitutes a debate. Here, I follow (Lauderdale and Herzog 2016, 14), who define a debate as set of contributions that share the same title, made on the same day, with a minimum of five speakers. For the estimation, speeches are concatenated per speaker for each debate, leaving us with 764,828 texts across 71,501 debates.

A comparative assessment of dimensionality

Like all text scaling models (or even simple cluster analyses or multidimensional scaling techniques), `Wordfish` tries to estimate a lower-dimensional, and simpler representation of the text data. The algorithm however estimates only one dimension, and, we thus have to be able to reasonably assume that this particular axis represents the main angle of "conflict" between actors. We then have to assume that variation in word usage is truly associated with the underlying latent dimension of conflict, rather than with simple topic variance. This raises a problem: the estimated level of polarization may not be based on ideology-related divergence, but instead on variation associated with the topic on which legislators speak. Rather than one, the debates involve multiple axes of conflict. The advantage of the two approaches outlined above is their ability to limit such topic-related variation. We can demonstrate this by considering the percentage of variation accounted for by the first axis from applying CA to the same data, which recovers multiple dimensions (Lowe 2013). This statistic is summarized for both approaches in Figure 1 below, which also includes a measure for baseline comparison that simply combines speeches for each legislator across all debates within sessions.

In the baseline approach, the principal axis of variation explains 4.7 percent on average.[14] The first strategy—which involves preselecting speeches based on a dictionary and machine classifier step—yields a significant improvement. When we only retain speeches that have a 75 percent or higher probability of falling in the economy category, this approach produces models where the first dimension explains 5.57 percent. When we increase the threshold to 99 percent, this rises to 8.09 percent.[15] When we reduce dimensionality further, by scaling legislators within debates (i.e. the `Wordfish` estimates from the first stage of `Wordshoal`, prior to factor analysis), we see a

[12] This approach is the closest approximation—in theoretical terms—to the CMP project, and to that of Slapin and Proksch (2008). A collection of an individual's speeches on a dimension in effect represents her "manifesto", that can subsequently be scaled.

[13] To provide an intuitive example of the second stage of the estimation procedure, imagine a set of three legislators (A, B, and C) who speak in three debates (D_1, D_2, D_3). In D_1, the `Wordfish` estimation places the legislators at: $A = 0.8$, $B = 0.5$, and $C = -0.3$. For D_2, the polarity is inverted, with $A = -0.8$, $B = 0.5$, and $C = 0.3$. This may happen because the debate-level scales could run left–right or right–left (because the model does not fix the polarity). The factor analysis will identify that the order of the three legislators is generally the same across the debates, and infer that the β coefficient relating the debate scale to the general scale is +1 (for example) in the former cases and −1 in the latter cases. Now imagine D_3, with the following scores: $A = 0.6$, $B = -0.3$, and $C = 0.1$. This debate is not very well correlated with the others. If most debates look more like D_1 and D_2, the factor analysis will estimate a weak loading (β) on that debate (i.e. it will contribute less to the *overall* latent dimension estimated from the debates). Conversely, if most debates look like D_3, the general dimension will look like this debate, and D_1 and D_2 will have small βs.

[14] The average correlation across the sessions between the first-dimension CA estimates and the `Wordfish` estimates is 0.94.

[15] The average correlation with the CA estimates is 0.92 and 0.85 respectively for these two approaches.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

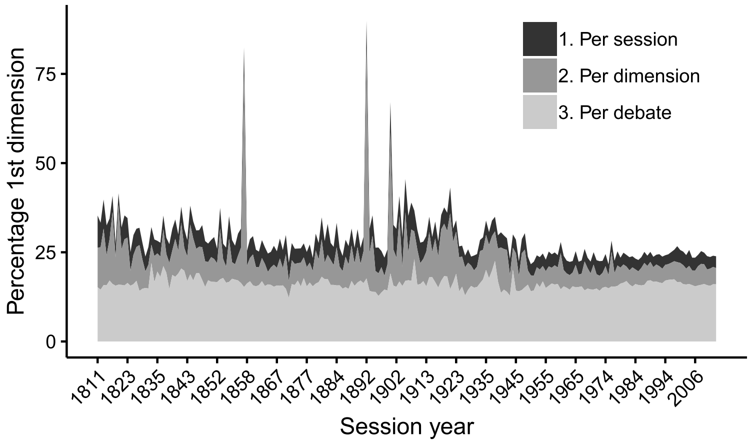

Figure 1. Stacked distributions of the perc. of variation explained by the first axis recovered by CA. “1. Per session” = speeches aggregated per session for each legislator; “2. Per dimension” = dimension-related speeches preselected using a classifier and subsequently aggregated for each speaker for each session; “3 Per debate”: speeches are aggregated by legislator for each debate and the model is estimated at the debate level.

significant improvement: here, the variation explained by the first axis rises to 16.26 percent on average.[16]

From ideal points to polarization

Based on the above, I conclude that `Wordshoal` is best able to recover a meaningful dimension of conflict. Subsequently, I measure polarization by “dummying out” the changes in the relative placement of legislators from session to session.[17] Specifically, I measure polarization as *the number of legislators of the right-most party that falls within the range of the distribution of legislator ideal points of the leftmost party, as a proportion of parliament*. To evaluate and reduce the effect of outliers, I also implement this second measure while only retaining legislators for the leftmost party whose position falls below the 95th percentile; and for the right those that fall above

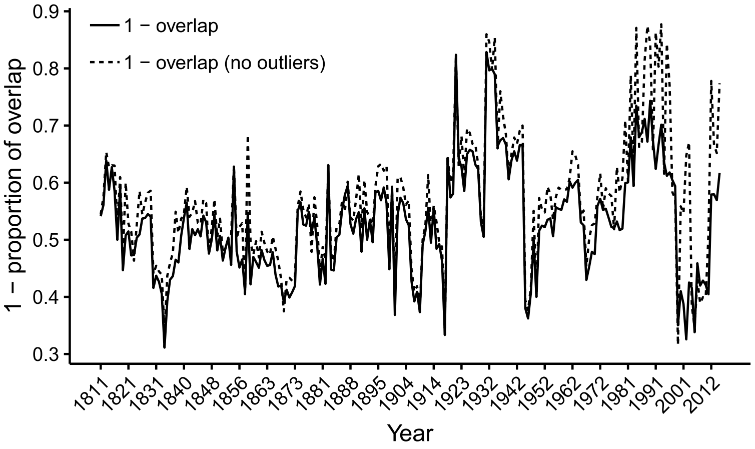

Figure 2. `Wordshoal`-based polarization estimates for 1811–2015.

[16] The correlation with the CA scores is 0.95. From the graphs, another property of the debate-level scaling approach becomes apparent: there is markedly less variability in the percentage of variation explained by the first axis. This is not surprising, as we may expect some upward, secular trend over time in the number of issues that parliament needs to legislate on and therefore discuss. This is especially the case after WWII, with the introduction and expansion of the welfare state.

[17] As the estimates are standardized, with fixed bounds at the extremes of the political spectrum, we are unable to compare the distance between the distributions of ideal points of parties from session to session, as these measures will not meaningfully reflect temporal changes.

the 5th percentile.[18] In both cases, I compute the polarization score by subtracting the proportion obtained from one. Figure 2 below plots the results.[19]

The measure thus captures the consistency with which MPs *fall within their party label across multiple policy issues*. A score of "1" represents perfect polarization, with zero overlap between the main parties on the right and left of the political spectrum.

5 Supervised Models: Machine Classification

We will now move beyond the shortcomings of unsupervised methods, and turn to a supervised model instead.[20] Specifically, I build on a novel machine classification approach (cf. Peterson and Spirling 2018), which ensures that we model variation related to the quantity of interest: *polarization*.

Polarization as partisan language

The classification accuracy approach of measuring polarization is based on a simple assumption of how language is generated that is similar to that of `Wordfish`. Language use reveals partisanship. In the US Congress, Republicans refer to "death taxes" and "illegal aliens", while Democrats will speak of the same issues using phrases such as "estate taxes" and "undocumented workers" (cf. Gentzkow, Shapiro, and Taddy 2016). In the UK, Conservatives might refer to a cut in benefits as "reducing dependence", while Labour may speak of a "benefits squeeze".

A key difference between unsupervised models and this supervised classification approach is that we introduce extra information—the party label. We therefore no longer have to limit the set of speeches that we feed into the model of political conflict *a priori*, as the party label is used to "pin down" the main axis of conflict. Instead of extracting one latent dimension, this *supervised* approach uses algorithms to identify the features (i.e. terms) associated with a particular party affiliation from labeled political speeches. These features, identified in the context of the complete body of texts to which the model is fitted, help us identify how partisan a particular speech (on the economy, foreign policy, welfare, etc.) is in relation to the corpus of speeches by that legislator's party.

The trained model "knows" how members of party A typically speak—it has "learned" the features of that party's language—and estimates the probability of an individual belonging to that party A for each speech that we "ask" it to predict.[21] As a basic intuition, a polarized parliament consists of groups that choose to use very distinct language; and an unpolarized legislature includes MPs who are linguistically proximate to members of their own party. Style, sub-topic, and other semantic differences are used strategically by legislators to make a point. The level to which this accords with a particular "party label" as predicted by a trained model thus reveals the degree of partisanship of the member.

This approach is particularly well-suited to high-dimensional data because we avoid the problem of issue space altogether. Disagreement is instead reduced to one dimension: language use. This broadens the concept of "ideology" as it is usually defined in the literature (see also Gentzkow, Shapiro, and Taddy 2016; Peterson and Spirling 2018). However, it can be seen as an efficient and appropriate approach if we accept the assumption that all—or at least a majority—of an MP's linguistic choices are informed by political considerations.

[18] The correlation between the two measures is 0.90 (at $p < 0.001$). The effect of outliers thus seems to be relatively limited.

[19] Here, I retain only Tories and Whigs prior to November 1922, and Conservatives and Labour thereafter.

[20] An additional problem with the unsupervised approach is the large error terms that occur when attempting to reduce sparsity of the word frequency matrices (WFMs). See section B1 in the on-line Appendix for a detailed discussion.

[21] In contrast to the `Wordfish` model, the placement of an individual legislator is not solely determined by word weights that are a function of the relative frequency with which terms are used within a debate; rather, the degree to which a word "loads" on (or: is predictive of) an MP's placement is a function of how likely that term is to be used by the party in general, across the session.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Implementation

To measure legislator preferences and parliament-level polarization, I apply the SGD classifier algorithm with a log loss function and l2 regularization (Bottou 2004).[22] In simple terms, the classifier algorithm is fitted to a randomly selected sample of speeches (i.e. the "training set") to identify what words and phrases are associated with a particular "class" (here: the party label). Subsequently, the algorithm is used to predict the party label of a second sample of test data from that same year. The degree to which language can accurately predict the party label is a measure of polarization.

Following Peterson and Spirling (2018), I use k-fold stratified sampling from each dataset of yearly speeches. Each year is randomly partitioned into twentyfolds of equal size while retaining the balance of party labels. Subsequently, one of the k subsamples is reserved for testing and the remaining $k - 1$ folds for training. By cross-validating twentyfold, I obtain individual-level partisan scores for each legislator using probability estimates for each label. In other words: the probability of any individual speech belonging to one "class" or the other represents a legislator's "partyness" on that particular occasion. The mean probability of belonging to their party across all the individual's speeches for a time period t represents that legislator's partyness for that period t.[23]

In contrast to Peterson and Spirling, I include *all* parties in the estimation, rather than the Conservatives and Liberals/Labour alone.[24] As the model's predictions depend on the data on which it is trained the inclusion of other, smaller parties are bound to affect the estimates, especially since different parties are more invested in some debates than in others. For example, between 2013 and 2014, Scottish independence featured prominently on the House's agenda.[25,26,27]

Results

For the estimation, I again use the cleaned-up data outlined above, which contains only speeches longer than fifty words, limited to entries with successful party label matches, and excluding

[22] To evaluate the impact of the chosen classifier on our estimates, I also use a multinomial Naive Bayes (NB) algorithm (Maron and Kuhns 1960). The SGD algorithm will serve as the main application. Results for the NB classifier are reported in Appendix E in the on-line supplementary materials.

[23] Also in line with Peterson and Spirling (2018), to account for differences in the length of speeches, and for important and common words, I apply the TF-IDF transformation to the WFMs (Manning, Raghavan, and Schütze 2008). This transformation up-weights words proportional to the number of times they appear in the document, and is offset by the frequency of the term across the corpus. Further, I also include a measure of uncertainty of the predictions across the folds. As this split is random, and the accuracy of the fitted model depends on the composition of the test sample, we can construct a measure of confidence on the basis of the set of accuracy estimates. Here, I use a bootstrap procedure ($n = 10,000$) to generate confidence intervals for the aggregate score per session.

[24] I do maintain a "hard" lower limit that excludes parties that contributed less than 100 speeches in a session. The reason for using an absolute rather than a relative (i.e. percentage) cut-off is the extreme dominance of the largest two parties (Liberals and Tories prior to 1922, and Labour and Tories thereafter).

[25] These include the debate on the Scottish Independence Referendum Bill, passed by the Parliament on 14 November 2013 (Royal Assent on 17 December 2013), and discussions on the Devolution Bill in light of the September 2014 independence referendum, as well as on the referendum itself. Excluding the Scottish National Party (SNP) from the model would bias the results for these discussions, and fail to include information on the degree to which legislators from the mainstream parties share the views of their SNP colleagues. Similarly, one can imagine that a model would produce biased results if it excludes the Irish Home Rulers in the 1880s, or fails to include *both* Liberals and Labour at a time when these two parties vied for the position of the main "second party" in the early 1900s.

[26] A second difference between my approach and that of Peterson and Spirling is that I conduct a "rough" grid search to tune the α hyperparameter, varying the α between $1e - 4$ en $1e - 7$, and selecting different levels for this parameter based on model performance evaluated by the classifier's accuracy. In practice, the value of α depends on the kinds of sample weights that are used (see next section).

[27] All algorithms were run on the Harvard-MIT Data Center Research Computing Environment.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Table 3. Validation scores.

Validity category	Test	ML	Wordshoal
1. Face validity	1.1 General test	✓	✓/✗
	1.2 Detailed test	✓	✗
2. Convergent validity	2.1 Session-level estimates	✓	✓
	2.2 Individual estimates	✓	✓
3. Construct validity	3.1 Between-session consistency	✓	✗
	3.2 Individual-level distribution	✓	✓/✗
	3.3 Explanatory power of party	✓	✗

procedural phrases. I implement SGD with class weights to balance between parties.[28,29] What drives our estimates, shown in Figure 4? As we are fitting a predictive model, a reasonable assumption would be that individuals who speak more—frontbenchers—have higher leverage on our predictions. We can verify this claim by considering the association between the status of speakers and the degree to which they are partisan—i.e. the probability of their party label. Here, I take the individual-level estimates of speeches for each session for the incumbent party and run a binary logistic regression to regress individual-level positions on a dichotomous response variable that captures frontbench status (frontbench = 1; nonfrontbench = 0).[30] This produces a log odds of 10.1 ($p < 0.001$). A (one-sided) t-test also shows that there is a statistically significant difference in mean polarization between the sample of frontbench and nonfrontbench MPs, with the former being higher.[31] There thus is evidence to suggest that our estimates in part reflect the proximity of a party's members to their ministerial team.

6 Validation Exercise

This section turns to validation exercise, the purpose of which is to demonstrate the usefulness of the validation framework, and to identify the estimation procedure that generates a measure that best maps onto our concept of polarization. From the implementations above, two candidates emerged: (i) a measure of overlap between parties' ideal points from an unsupervised `Wordshoal` model; and (ii) the predictive accuracy from a machine classifier that is trained on labeled speeches from different parties, with weights to account for imbalances between parties. The estimates from these approaches do not correspond closely ($\rho = 0.13$ at $p = 0.04$), making a validation exercise even more pertinent. Table 3 above shows the performance of the estimates in accordance with the validation framework.

An important quality of *meaningful* estimates from text-based measures of polarization is that they should correspond with our expectations of how the measure develops over time. As part of this face validity criterion, the **general test (1.1)** considers the stability of estimates within

[28] The weights are defined as $\frac{n_{\text{total}}}{p * n_p}$, where n_{total} is the total number of speeches, p is the number of unique parties, and n_p = number of speeches of party p.

[29] Note that there is little difference between the implementation with and without party class weights at the aggregate level (i.e. the classifier level). The correlation between the aggregate-level and individual-level accuracy scores of both implementations are 1.00 and 0.92 respectively. This likely is a function of the fact that the parties are already well balanced in the sample: Until November 1922, the Tories make approximately 46 percent of speeches in the sample; whereas the Liberals account for 54 percent. From 1922 until 2015, the discrepancy is even lower, with the Conservatives at 52 percent, and Labour at 48.

[30] Frontbench status is defined as whether the MP is a government minister or not. The focus on members of the incumbent party is motivated by two limitations of the data. First, the raw data provided by Political Mashup only include information on MPs' membership of the government from the second session of the 37th parliament onwards, starting on 3 November 1936. Second, the labels included in the data exclusively indicate members of the government; they do not mark MPs as being in the shadow Cabinet or not. I can therefore not run a similar test for the opposition party. However, we can still have confidence given that there is alternation in government status between parties and there is no reason to assume that party leaders would systematically shift position after switching from Shadow Cabinet to incumbency status.

[31] Frontbench MPs show a 9.3 percent higher probability of being in their party.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

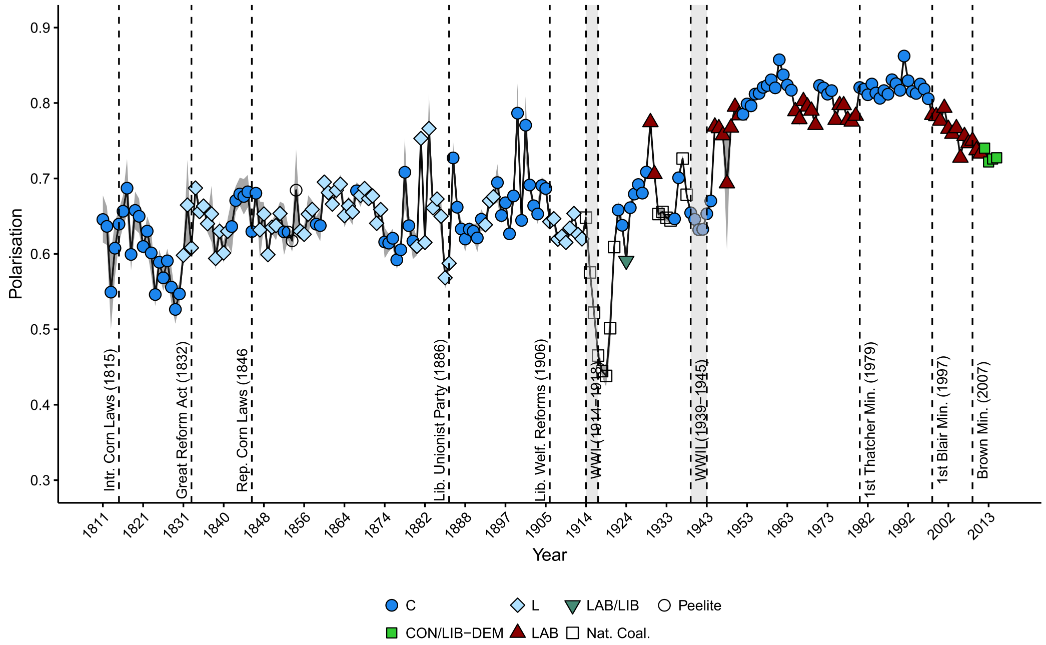

Figure 3. Illustrated plot of the accuracy of the SGD classifier for each session of the UK House of Commons for 1811–2015 (details described in text).

parliaments. One element of immediate concern is the implausibly high level of variability shown in the estimates within parliaments for the Wordshoal-based measure (Figure 2). The changes are dramatic, suggesting an almost random pattern of switches between high and low levels of polarization that are uncorrelated between sessions. By contrast, the measures derived from the classifier yield more stable results (Figure 3), with relatively high correspondence between sessions within a parliamentary term that appear to map onto a "stable political space".

At a more granular level, the **detailed test (1.2)** similarly suggests that the machine classifier is able to produce estimates that correspond to our *a priori* expectations in a way that Wordshoal cannot. Figure 3 shows that this first measure corresponds well with important historically identifiable outliers in polarization (Table 2). Polarization grows in the wake of the Corn Laws

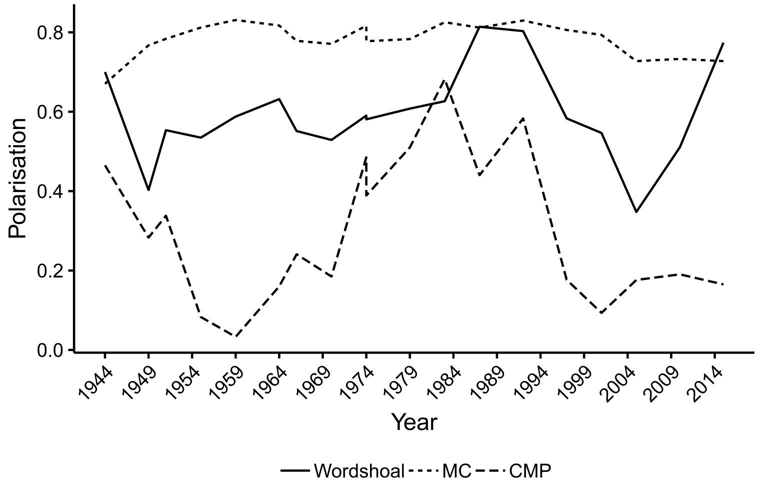

Figure 4. Comparison of estimates with Comparative Manifesto Project rile scores. Measures of polarization based on Wordshoal, the SGD classifier, and the rile scores respectively, for each election year covered by the CMP data.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

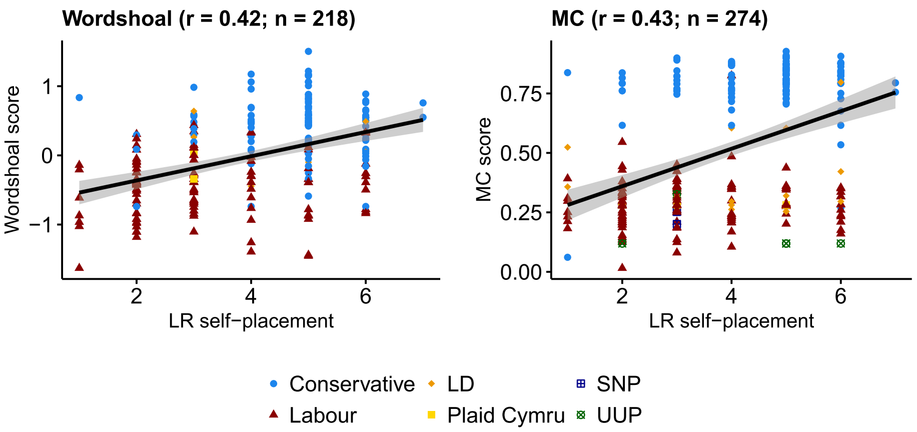

Figure 5. Scatter plots and regression lines for the association between the left–right self-placement of surveyed MPs in the 1992 wave of the British Candidate Survey (Norris and Lovenduski 1995) and the individual-level estimates of the `Wordshoal` and SGD implementations respectively.

(Table 2, id 1) and of the 1832 Reform Act (id 2), and is generally higher in the period after 1880 (ids 3 & 4). The formation of the Liberal Unionist party in 1886 appears to mark the start of a period during which members did not fall consistently within their party label, which explains the rather dramatic drop in that year. Although they generally agreed with the Conservatives on Ireland, they were still classed as "Liberals" (at least for part of the time), which makes aggregate polarization look very low.

After 1906, at the start of the Liberal Welfare Reforms, we see greater polarization over these "controversial" new policies (id 4). As one would expect, we also see less conflict between members of different parties during WWI and WWII (ids 6 & 8), and during the 1923 MacDonald Ministry (id 7), with levels picking up further after the 1945 landslide Labour election victory. Finally, while the Thatcher ministries of 1979–1990 appear to be highly polarized (id 9), a decline in polarization may be observed with the start of the Brown government (2001) and of the coming into office of the Conservative–Liberal Democrat coalition (2010) (id 10).[32]

Our second set of tests focuses on the convergence of the estimates with an exogenous measure. First, to investigate the correspondence with **session-level estimates (2.1)**, comparable data are only available for the period after 1945, for which we can analyze the convergent validity of our estimates with the `rile` score (Laver and Budge 1992) of the parties based on the CMP data (Volkens *et al.* 2016).[33] Figure 5 above presents a visual comparison.[34] Here, the sessional score for the year *preceding* the election year is matched with the CMP scores.[35] These results paint a mixed picture. Both the classifier and the `Wordshoal` results bear a reasonable resemblance to the `rile`-based score (with the exception of 1960). The latter however seems to show better correspondence. This is not altogether surprising as the `Wordshoal` approach should be more sensitive to changes to the issues on the agenda (as the CMP's `rile` scores are too) than the machine classifier. As outlined in the previous sections, the former approach extracts a latent dimension from each debate and subsequently runs a Bayesian factor analysis to extract a score across these debates for each MP. As new issues enter the parliamentary arena, we can expect

[32] The downward outlier of 1948 is a consequence of the fact that this was a short session (14 September–25 October 1948) with a correspondingly low number of substantive debates over which there could be a division of opinion. With little data to train on, the SGD algorithm will have more trouble predicting party labels on the basis of speech, as it should in this context.

[33] I take the absolute difference between the Labour Party's and the Conservatives' `rile` score divided by 100 as a measure of polarization.

[34] Unfortunately, the data is only available for every *election*, and not for each session like the scores I develop. As such, we have too little data to consider correlations or other similarity measures.

[35] This is to ensure that we do not include legislators elected in the new parliament in our comparison.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

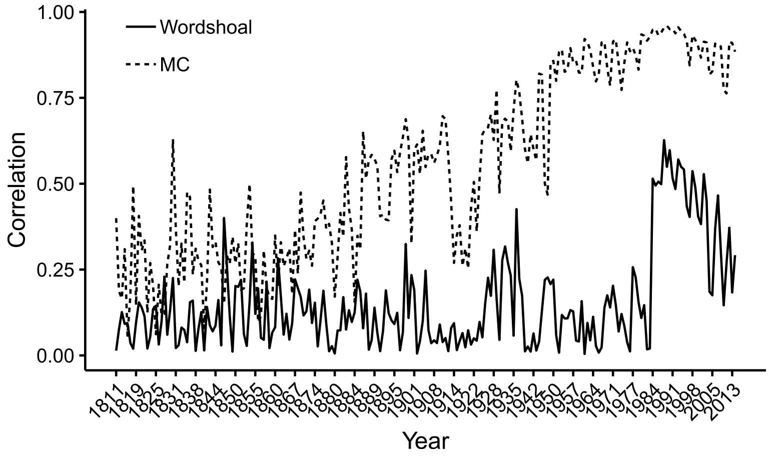

Figure 6. Correlation of individual-level estimates between session t and session $t + 1$ for all parliamentary sessions of the UK House of Commons from 1811 to 2015.

these positions to shift, affecting the aggregate measure of polarization. Conversely, although a dynamic whereby parliament needs to engage with new issues changes the features of the corpus on which the classifier is trained, this latter approach will not treat these as new dimensions. Debates on novel problems may make the classifier more or less accurate (depending on how divisive the issue is for parties), but it will not contribute as directly to changing the position of individual legislators as would happen in the Wordshoal approach.

A more promising level for comparison is that of **individual estimates (2.2)**, for which we can rely on data from the 1992 wave of the British Candidate Survey (BCS) (Norris and Lovenduski 1995).[36] The BCS asks respondents to rank themselves on a seven-point ordinal left-to-right scale. I match these records from the 1992 wave—the availability of which is, of course, limited by response rates—with my own MP-level estimates (taking their maximum prediction value). I do so for the first session of the 1992–1997 parliament, as this is closest (in time) to when MPs responded to the survey (i.e. in 1991). The results (Figure 5) show that the estimates correlate most strongly with the classifier results ($\rho = 0.43$), but followed closely by Wordshoal (at $\rho = 0.42$). Naturally, we cannot extrapolate to the full sample, but these results are nevertheless encouraging. Specifically, they suggest that the machine classifier is, similar to Wordshoal, able to produce results that bear a close relationship to *the position that legislators give themselves on a left-to-right scale*.

Examining the stability of the estimates over time (**between-session consistency test (3.1)**), allows us to establish whether the estimated positions reflect long-held political views of legislators, or, alternatively, represent issue-specific divergences. Such stability is crucial when it is our intention to use a polarization measure in a substantive application, i.e. to test hypotheses that relate to political phenomena across extended time periods. It ensures that the measure is comparable over time, that is, that it relates to the same construct rather than to issue-specific divergences. To assess between-session consistency (or: stability), we consider the correlation from one year to the next for legislators in each parliament. Figure 6 plots these correlations for all sessions for the MC and the Wordshoal estimates. For the former, the average correlation across the sample is 0.55, and there is a steadily upward trend, with the highest level of session-to-session consistency in the 1980s. The correlations for Wordshoal have a mean of 0.15, and range between $4.3e - 3$ and 0.63. This latter result raises some issues for the unsupervised scaling technique. We would expect Wordshoal to be *less* consistent between sessions (for reasons outlined above).

[36] Since 1997, data such as constituency references were taken out to ensure anonymity, so I cannot use later iterations for comparison.

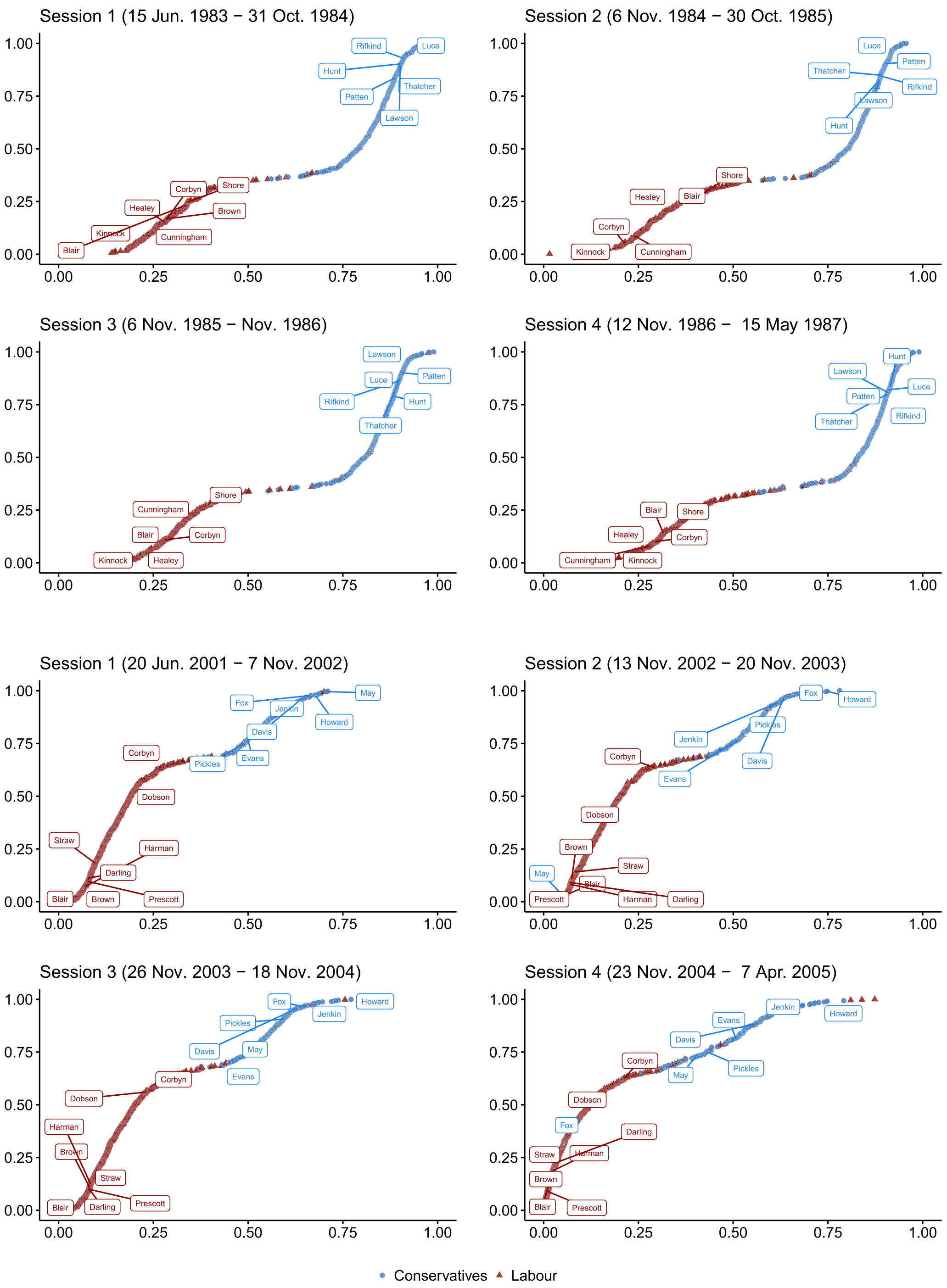

Figure 7. Empirical Cumulative Distribution Functions of the individual-level class probabilities for legislators obtained with the supervised machine classifier for all sessions of the 1983–1987 and the 2001–2005 parliamentary terms.

However, we should still expect legislators to be somewhat consistent in their positions across different issues that make it to the agenda. Again, it appears that the machine classifier is better able to capture the position of individual legislators over time.

Further, I analyze whether the **individual-level distribution (3.2)** of estimates shows a clear division of legislators between the main parties, and whether the placement of key legislators on either extreme matches our expectations. Here, I consider the 49th parliament under the second ministry of Margaret Thatcher (1983–1987), and Tony Blair's government of 2001–2005 (53rd

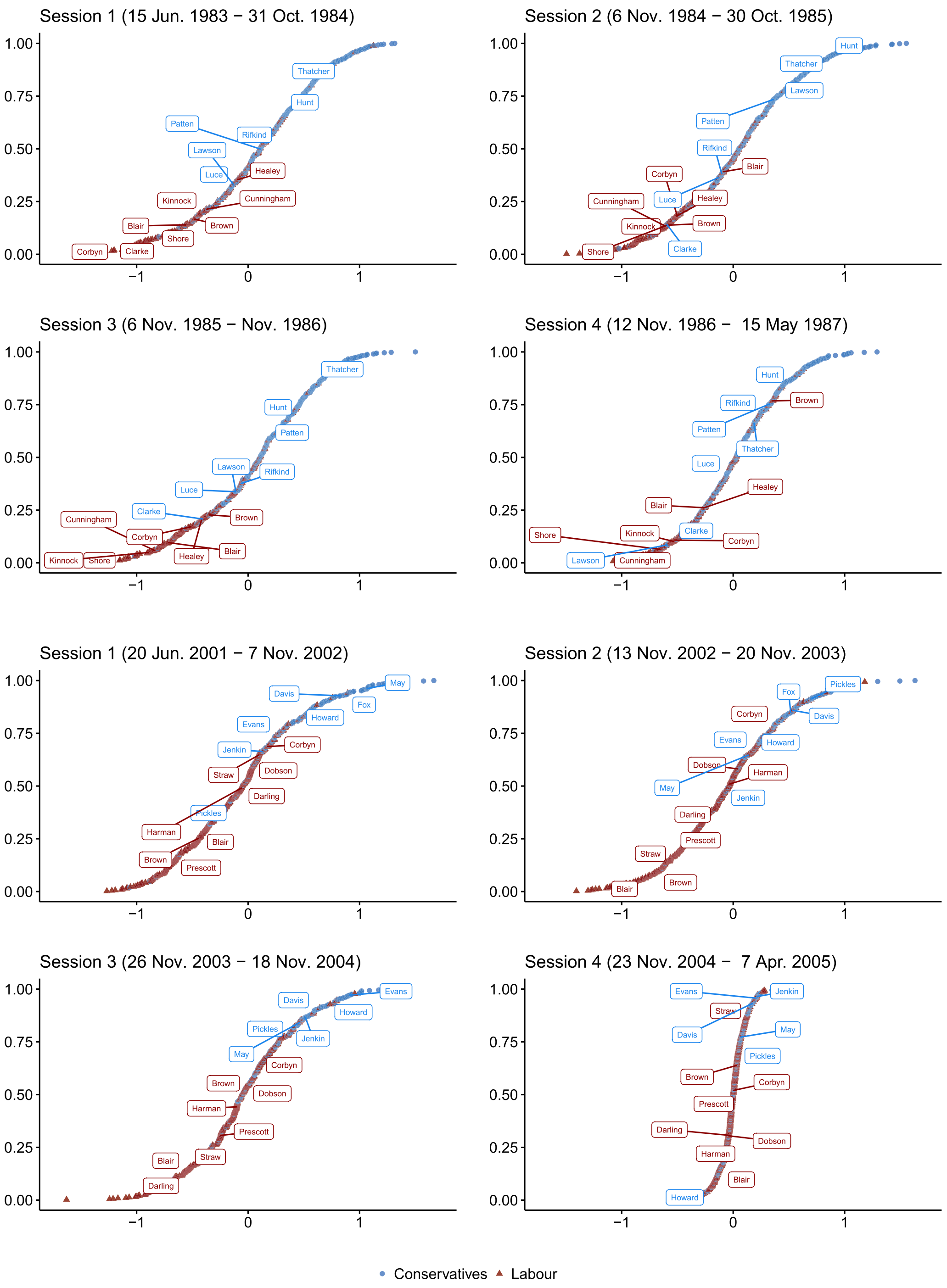

Figure 8. Empirical Cumulative Distribution Functions of the individual-level class probabilities for legislators obtained with the `Wordshoal` algorithm for all sessions of the 1983–1987 and the 2001–2005 parliamentary terms.

parliament). Figure 7 plots the ECDFs of the individual-level estimates of the classifier accuracy approach for each session (four for each parliament). An "individual estimate" is the mean accuracy of all speeches an MP made in a session. Figure 8 plots the ECDFs of the legislator-level estimates obtained with `Wordshoal`. Here, the unit of observation is the factor score across all debates in which an MP participated, as described earlier in this paper.

An important "visual" test of the plausibility of the two measures is whether the ECDF clearly classifies members as belonging to one party. We should not expect perfect separation in every

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

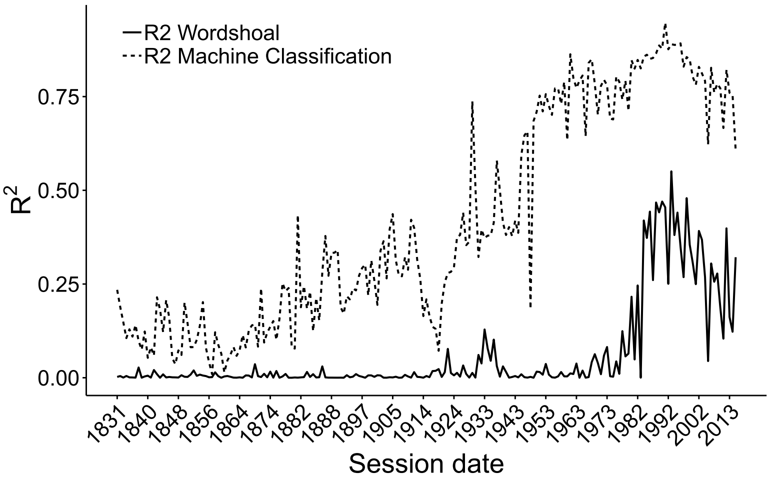

Figure 9. R^2 from regressing MP-level estimates on party label.

parliament—after all, polarization varies over time—but, a complete overlap of parties seems equally implausible. The classifier estimates show a clear division. Further, key politicians in the Shadow Cabinet and the Cabinet are placed further out in the tails, as one would expect. For example, from 1983 to 1987, PM Thatcher and the Chancellor of the Exchequer Nigel Lawson (as well as other key figures) are clearly out in the tails (Figure 7). Conversely, their counterparts for the 2001–2005 parliament, Tony Blair and Gordon Brown, are clearly placed on the extreme left of the spectrum.[37]

Second, the `Wordshoal`-based estimates show relatively good separation of parties, but still inspire less confidence. Figure 8 shows that there is a considerable degree of overlap between members of different parties. Although I have no *a priori* expectations as to how concentrated the parties should be, one would at the very least expect stronger clustering of parties, and the separation should certainly outperform an "at-random" distribution of estimates.

Finally, the validation framework prescribes an investigation of the **explanatory power of the party (3.3)**: do these text-based approaches capture anything beyond simple government–opposition dynamics? To evaluate this question, we take a simple linear model for each session where we regress individual scores on their party's mean position. These measures are plotted in Figure 9. It is clear that the SGD algorithm is not simply capturing party affiliation. There is a relatively strong correspondence between party position and label, but the levels of the R^2 show that some unexplained variation remains. It has a mean of 0.39 across the sample, and a minimum and maximum of 0.01 and 0.95 respectively.[38] We obtain different results for the `Wordshoal` estimates. Here the range is $[1.09e - 8, 0.55]$ with a mean of 0.06. These values are implausibly low and reaffirm our findings above that the division between parties in this approach less clearly reflects political affiliation.

[37] The position of Theresa May in these plots is—albeit perhaps anecdotal—also somewhat revealing of what the individual-level estimates of the machine-learning approach are capturing. We would expect May to be to the extreme right, given her position as Chairman of the Conservative Party in 2002–2003, and her role as Shadow Secretary of State for the Family in 2004–2005. Both roles make her part of the party frontbenches, and we would therefore expect May to use language that identifies her closely as a member of the Conservative Party. However, it is plausible that in her latter role, she would have had to use vocabulary that is typically used by Labour politicians. For the 2002–2003 period, this indeed seems to be the case: speeches made by Theresa May for this session include remarks made for international women's day, and the top twenty words used include "women", "equality", and "education". In the 4th session of the 53rd Parliament, 78 percent of May's speeches relate to the topic of family justice, and words such as "children", "family", "parents", and "services" are among those used most frequently.

[38] When limiting the sample to Tories/Conservatives, Liberals, and Labour, these figures are 0.37, 0.01, and 0.95 respectively. Figure 9 shows estimates for a sample that includes all parties.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

7 Discussion and Conclusion

The use of speech data to inform our understanding of parliamentary polarization is still in its infancy, and presents researchers with significant challenges. Using over 6.2 million speech records from the UK House of Commons, this paper has outlined a widely applicable framework for validating such text-based measures of polarization, which consists of clear and manageable tests that researchers can rely on. I have demonstrated the framework's usefulness in an application to an unsupervised scaling technique (`Wordshoal`) (Lauderdale and Herzog 2016), and a novel *supervised* machine classifier approach (Peterson and Spirling 2018).

These applications suggest that unsupervised (scaling) approaches that do not incorporate information about party affiliation fail to produce estimates that map onto a clear and temporally comparable political space. Conversely, a simple machine classifier strategy that puts party information front and center produces estimates that show a high degree of face-, construct-, and convergent validity. This finding becomes most strongly apparent when considering the parts of the validation framework that consider the individual-level estimates. In contrast to `Wordshoal`, the machine classifier approach produces measures of legislator positions that show much greater stability over time (test 3.1), that separate parties well and place key individuals correctly in the political space (test 3.2), and that correspond to MPs' self-reported ideological positions (test 2.2). The machine classification method is thus particularly suited to researchers who seek to apply text-based measures of polarization in substantive applications, and in particular in studies that focus on a long time period.

Many of the problems of supervised approaches such as `Wordfish` (and its sibling, `Wordshoal`) seem to stem from the fact that they limit themselves to estimating one, latent dimension, which is the axis that accounts for the greatest amount of variation in word use. There is no guarantee that this particular axis corresponds to the dimension of party conflict that we are interested in. The strength (and weakness) of the classification approach is that we can pin down the target that we want to capture. This gives us estimates that we can reasonably assume (and, through validation can be shown) to be related to conflict between political parties. As in many cases we have information on the party to which individuals belong, we should in the estimation of polarization rely on this superior method.

This conclusion is not simply an artifact of something unique to the UK data. An additional validation exercise, for which I use the same data from the Irish Dáil and the US Senate from Lauderdale and Herzog (2016), reveals that the supervised, machine classifier approach also performs well in these contexts, at least when it comes to identifying opposition and government members (see Appendix F in the on-line supplementary material for a detailed comparison). For example, while the R-squared from regressing estimated positions on party labels is lower for the machine classification approach in both the Dáil and the US Senate, the ECDFs show that in both cases key legislators are placed where we would expect them to be on the distribution. In addition, the estimates for the US Senate correlate well with exogenously created measures.[39]

Researchers that do wish to rely on unsupervised scaling techniques should think more carefully about limiting the lexicon to which they apply scaling techniques to the area of

[39] Whereas Lauderdale and Herzog (2016) report a correlation of Bonica (2014) career CFscores (based on campaign donations) with their speech scores of $\rho = 0.55$, the machine classification achieves a correlation of $\rho = 0.92$ across the whole sample of the 104th–113th Congress (taking the prediction of the being Republican as the score for the machine classification model). This however appears to be to some degree driven by the strong distinction between Republicans and Democrats identified by the supervised model: the latter party's members are concentrated at the extreme right of the distribution, whereas democrats have a very low probability of belonging to this party label. More encouragingly, I also find a high correlation between Gov.track ideology scores for 2016 (based on bill and resolution co-sponsorship, cf. GovTrack.us 2013) and my machine classification estimates. Although imperfect given the temporal mismatch, for the 68 Senators who continue in the 112th Congress, I find a correlation between these ideology scores and the estimates from the supervised model of $\rho = 0.94$ for the complete sample, of $\rho = 0.41$ for Republicans, and, even more encouraging, of $\rho = 0.66$ for Democrats. For the 60 Senators who continued in 2017, these correlations are at 0.91, 0.33, and 0.58 respectively.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

substantive interest (or "dimension") that they seek to analyze. I have suggested two strategies to do so. First, in the parliamentary context one can sift out procedural terms using an "endogenous" dictionary approach, i.e. using records of the parliament's own procedures (see the on-line supplementary materials). Second, we can reduce the dimensionality of the semantic space by: (i) lowering the level of analysis to individual debates; and (ii) applying a two-step dictionary approach using dictionaries and semantic classifier algorithms to select relevant speeches. Even while applying such selection techniques, however, it appears that the supervised model outperforms the unsupervised variant.

Two areas for improvement stand out. First, the strength of the classification accuracy approach lies in ignoring dimensionality. In so doing however, we sacrifice our ability to make substantive claims about the drivers of conflict. When we say that the House of Commons is polarized based on the ability of language use to predict party affiliation, what is the axis of disagreement? The economy, security issues, or, perhaps, foreign policy? A possible solution to this problem is to first preselect speeches on a specific dimension—for example by using my two-step dictionary and classification approach—and subsequently apply the classifier. This would allow researchers to analyze political disagreement on a more granular level.

Second, while we have a good appreciation of how institutional dynamics affect roll-call votes (Spirling and McLean 2007; Hix and Noury 2010), we do not have a comparable level of understanding of how they impact text-based estimation. The degree to which legislators engage in debate is subject to both cross- and within-country variation (cf. Benedetto and Hix 2007; Kam 2009; Eggers and Spirling 2014; Proksch and Slapin 2015)—dynamics which our models can and should incorporate. A comprehensive machine classification approach to measuring polarization would distinguish appropriate weights to account for individual- and system-level characteristics, as well as cross-temporal dynamics such as the safety of the MP's seat and exogenous shocks. I leave such and other improvements for future work.

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2019.2.

References

Adcock, R., and D. Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Benedetto, G., and S. Hix. 2007. "The Rejected, the Ejected, and the Dejected: Explaining Government Rebels in the 2001–2005 British House of Commons." *Comparative Political Studies* 40(7):755–781.

Binder, S. A. 1996. "The Partisan Basis of Procedural Choice: Allocating Parliamentary Rights in the House, 1789–1990." *The American Political Science Review* 90(1):8–20.

Bonica, A. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386.

Bottou, L. 2004. "Stochastic Learning." In *Advanced Lectures on Machine Learning*, edited by O. Bousquet, U. von Luxburg, and G. Rätsch, 146–168. Berlin and Heidelberg: Springer.

Carrubba, C. J., M. Gabel, and S. Hug. 2008. "Legislative Voting Behavior, Seen and Unseen: A Theory of Roll-call Vote Selection." *Legislative Studies Quarterly* 33(4):543–572.

Carrubba, C. J. et al. 2006. "Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-call Vote Analysis." *British Journal of Political Science* 36(4):691–704.

Cox, G. W. 1987. *The Efficient Secret: The Cabinet and the Development of Political Parties*. Cambridge: Cambridge University Press.

Diermeier, D., and R. Vlaicu. 2011. "Parties, Coalitions, and the Internal Organization of Legislatures." *American Political Science Review* 105(2):359–380.

Eggers, A. C., and A. Spirling. 2014. "Electoral Security as a Determinant of Legislator Activity, 1832–1918: New Data and Methods for Analyzing British Political Development." *Legislative Studies Quarterly* 39(4):593–620.

Gentzkow, M., J. M. Shapiro, and M. Taddy. 2016. "Measuring Polarization in High-dimensional Data: Method and Application to Congressional Speech." Working Paper.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press

Goet, N. D. 2018. “Replication Data for: Measuring Polarisation with Text Analysis - Evidence from the UK House of Commons, 1811–2015.” https://doi.org/10.7910/DVN/HHOIA4, Harvard Dataverse, V1.

GovTrack.us. 2013. “Ideology Analysis of Members of Congress.” https://www.govtrack.us/about/analysis.

Greenacre, M. 2016. *Correspondence Analysis in Practice*. 3rd edn. Boca Raton, FL: Chapman & Hall/CRC Press.

Grimmer, J. 2013. *Representational Style in Congress: What Legislators Say and Why it Matters*. Cambridge: Cambridge University Press.

Grimmer, J., and B. M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–297.

Herzog, A., and K. Benoit. 2015. “The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent During Economic Crisis.” *The Journal of Politics* 77(4):1157–1175.

Hix, S., and A. Noury. 2010. “Scaling the Commons: Using MPs’ Left–right Self-placement and Voting Divisions to Map the British Parliament, 1997–2005.” Paper prepared for presentation at the annual meeting of the American Political Science Association in Washington, DC, September 2–5, 2010.

Hug, S. 2010. “Selection Effects in Roll Call Votes.” *British Journal of Political Science* 40(1):225–235.

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

Lauderdale, B. E., and A. Herzog. 2016. “Measuring Political Positions from Legislative Speech.” *Political Analysis* 24(3):374–394.

Laver, M., K. Benoit, and J. Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 97(2):311–331.

Laver, M., and I. Budge. 1992. “Measuring Policy Distances and Modelling Coalition Formation.” In *Party Policy and Government Coalitions*, edited by M. Laver and I. Budge, 15–40. Basingstoke: Macmillan.

Lowe, W. E. M. 2013. “There’s (Basically) Only One Way To Do It: Some Unifying Theory for Text Scaling Models.” Paper prepared for the American Political Science Association Meeting, Chicago, September 2013.

Lowe, W. E. M., and K. Benoit. 2013. “Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark.” *Political Analysis* 21(3):298–313.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Maron, M. E., and J. L. Kuhns. 1960. “On Relevance, Probabilistic Indexing and Information Retrieval.” *Journal of the ACM* 7(3):216–244.

McLean, I., and C. Bustani. 1999. “Irish Potatoes and British Politics: Interests, Ideology, Heresthetic and the Repeal of the Corn Laws.” *Political Studies* 47(5):817–836.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16(4):372–403.

Norris, P., and J. Lovenduski. 1995. *Political Recruitment: Gender, Race and Class in the British Parliament*. Cambridge: Cambridge University Press.

Peterson, A., and A. Spirling. 2018. “Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems.” *Political Analysis* 26(1):120–128.

Proksch, S.-O., and J. B. Slapin. 2014. “Words as Data: Content Analysis in Legislative Studies.” In *The Oxford Handbook of Legislative Studies*, edited by S. Martin, T. Saalfeld, and K. Strøm, 126–144. Oxford: Oxford University Press.

Proksch, S.-O., and J. B. Slapin. 2015. *The Politics of Parliamentary Debate: Parties, Rebels and Representation*. Cambridge: Cambridge University Press.

Pugh, M. 1982. *The Making of Modern British Politics, 1867–1939*. Oxford: Basil Blackwell Publisher Limited.

Pugh, M. 1999. *State & Society: A Social and Political History of Britain 1870–1999*. 2nd edn. New York: Oxford University Press.

Quinn, K. M. et al. 2010. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54(1):209–228.

Schwarz, D., D. Traber, and K. Benoit. 2017. “Estimating Intra-party Preferences: Comparing Speeches to Votes.” *Political Science Research and Methods* 5(2):379–396.

Slapin, J. B., and S.-O. Proksch. 2008. “A Scaling Model for Estimating Time-series Party Positions from Texts.” *American Journal of Political Science* 52(3):705–722.

Spirling, A. 2014. “British Political Development: A Research Agenda.” *Legislative Studies Quarterly* 39(4):435–437.

Spirling, A., and I. McLean. 2007. “Uk OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons.” *Political Analysis* 15(1):85–96.

Vandoren, P. M. 1990. “Can We Learn the Causes of Congressional Decisions from Roll-call Data? *Legislative Studies Quarterly* 15(3):311–340.

Volkens, A. et al. 2016. *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2016a*. With Werner, Annika. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.

https://doi.org/10.1017/pan.2019.2 Published online by Cambridge University Press