Measuring the Significance of Policy Outputs with Positive Unlabeled Learning

Radoslaw Zubek, Abhishek Dasgupta, David Doyle

Abstract

Identifying important policy outputs has long been of interest to political scientists. In this work, we propose a novel approach to the classification of policies. Instead of obtaining and aggregating expert evaluations of significance for a finite set of policy outputs, we use experts to identify a small set of significant outputs and then employ positive unlabeled (PU) learning to search for other similar examples in a large unlabeled set. We further propose to automate the first step by harvesting 'seed' sets of significant outputs from web data. We offer an application of the new approach by classifying over 9,000 government regulations in the United Kingdom. The obtained estimates are successfully validated against human experts, by forecasting web citations, and with a construct validity test.

Forthcoming in *American Political Science Review*.

In modern democracies, new policies are frequently introduced, but not all are significant in social and economic terms. Identifying and classifying important policy outputs has long been of major interest to political science. Much research relies on measures of policy significance. Evaluations of policy outputs are, in particular, crucial for empirical investigations into how 'reform productivity' is shaped by divided government (Mayhew 1991), veto players (Tsebelis 1999; Conley and Bekafigo 2010; Angelova et al. 2018) and agenda control (Döring 2001; Becher 2010). Various ratings of policy outputs have also underpinned research examining how political conflict and uncertainty shape the autonomy and discretion of bureaucratic actors (Huber and Shipan 2002; Junge, Koenig, and Luig 2015).

The conventional approach to measuring policy significance relies on evaluations by media reporters and policy scholars. Experts observe the policy environment and classify new outputs into those that are sufficiently notable and those that fall below some threshold of significance (Mayhew 1991; Tsebelis 1999; Conley and Bekafigo 2010; Angelova et al. 2018). Once obtained, individual expert rankings are combined into binary, categorical or continuous scores using some aggregation method (Howell et al. 2000). Clinton and Lapinski (2006) provide the most rigorous such aggregation method, using item-response theory to extract latent policy significance from expert scores.

In this work, we offer a proof of concept for a novel measurement approach. Instead of collecting and aggregating expert evaluations for a finite set of policy outputs, we propose to (i) use experts to identify a (relatively) small set of significant outputs and then to (ii) search for sufficiently similar examples in a large unlabeled set. We further propose to automate the first step by harvesting 'seed' sets of significant outputs from web data and to implement the second step by applying advances in computational semi-supervised learning. We offer an application of the new approach by classifying a specific type of policy output – over 9,000 government regulations in the United Kingdom. We evaluate our classification by examining correspondence with human expert ratings; forecasting web citations of significant regulations; and testing the general construct validity of our significance estimates.

Our approach has important advantages. It is less time-consuming and labor-intensive than conventional methods. By harvesting 'seed' sets from the web, we are able to choose experts who evaluate outputs from a distinct perspective and to control the temporal dimen-

sion of such evaluations through date-restricted searches. In our application, we search for contemporaneous citations by leading law firms, a rater group that gives particular emphasis to policies that change the status quo by a large margin. Using positive unlabeled learning, we train a computational model that achieves a true positive rate of 85% when predicting future web content, correctly classifies 70% of hand-coded expert ratings, and produces estimates with high construct validity. Our method is not without its limitations. As with any automated method, a trade-off exists between labeling expense and prediction accuracy, and our approach achieves moderate success in classifying more nuanced cases.

A NEW APPROACH

Modelling Positive Examples

The approach we propose starts from a set of examples of significant policy outputs identified by a group of domain experts. We assume that the probability of a domain expert (rater) j identifying some output i as significant ($y_{ij} = 1$) is a sigmoid function on $a_j(\theta_i - b_j)$ where θ_i is the latent significance of output i, a_j is the discrimination of rater j and b_j is the threshold of rater j beyond which she rates a policy output as significant (Eq. 1). We assume that θ denotes the true significance of a policy output, and if it were observable, all raters would agree on the ranking of outputs, even though they apply different thresholds and discrimination rates. Raters are also assumed to share a common understanding of what determines significance (cf. Howell et al. 2000, 303-304; Clinton and Lapinski 2006, 238).[1]

$$\Pr(y_{ij} = 1|\theta_i, a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{e^{a_j(\theta_i - b_j)} + 1} \tag{1}$$

Suppose now that we obtain a set of outputs which we know rater j considers significant. Suppose further that we obtain k such sets from different experts in a given domain. Taking the probability of 0.5 as a decision boundary for inclusion in individual raters' sets of significant examples, the resulting set P of such outputs can be viewed as containing significant policy outputs in two respects. First, because a_j, the rater discrimination can be

[1]The assumption of a single dimension for θ is warranted as long as one selects experts from a specific domain who view output significance from a similar perspective. Appendix E shows this assumption holds in our application. If experts use multiple dimensions of significance, Eq. (1) must be adapted to reflect the multidimensionality of parameters θ, a and b, and the interpretation of the set P becomes less intuitive.

assumed to be always positive, we know that $\theta_i > b_j$ must be true for at least one rater j. In simple terms, the obtained set P is a set of significant outputs in the sense that the latent significance of each output in P is higher than the threshold of at least one rater.

Second, if at least some outputs in the set P are considered significant by two or more raters, we can make further inferences about θ for subsets of P based on parameter n, where $n \in \{1, ..., k\}$ is the minimum number of raters who have rated a given output i as significant. More specifically, by selecting different values of n, we can vary the lower bound on the significance of outputs in terms of rater thresholds. Assuming rater thresholds such that $B = \{b_1, b_2, \ldots, b_k\}$ for k raters and assuming that $b_1 \leq b_2 \leq \cdots \leq b_k$, we know that $\theta_i > b_1, ..., b_n$ for every output i in the subset $P_n \subset P$. In conclusion, as one increases n and takes corresponding subsets P_n, one obtains sets of policy outputs that can be considered to be of higher significance in terms of relative rater thresholds.[2]

Harvesting Positives from Web Data

Examples of significant policy outputs can be obtained by directly surveying a group of domain experts (Warber, Ouyang, and Waterman 2018). In our work, we propose to obtain such examples using web citations. This approach has important advantages. First, a few billion internet users worldwide upload millions of posts every day on a myriad of issues including public policies. By posting content on-line, users convey an evaluation of what policy outputs they consider significant. These assessments are readily available and require little cost to obtain. Importantly, many web content contributors (e.g. banks, market analysts, law firms, business associations) can be regarded as specialized domain experts and research can take advantage of their propensity to freely share professional opinions.

Second, by harvesting citations of policy outputs from the web, we can formulate more specific assumptions about the primary concept of significance used by our domain experts. Previous work tends to define significance of policy outputs in terms of general 'notability' in the eyes of elite evaluators (Clinton and Lapinski 2006). With our approach, we can draw on the copious research on online journalism and web content marketing (see e.g. Schultz et

[2]Note that because raters cannot be assumed to have considered all outputs, if output i is selected by rater j but not rater k, we cannot conclude that $a_k(\theta_i - b_k) < 0$ and $\theta_i < b_k$. This implies that applying standard methods (such as IRT) for estimating parameters a_k, θ_i and b_k can be problematic.

al. 2013) to make better informed assumptions about *why* a specific category of experts may judge some policy outputs significant enough to cite them online.[3] Our approach thus allows us to choose expert raters who can be assumed to use a distinct concept of significance.[4]

Last but not least, by searching for citations made on the web, we are able to control the temporal dimension of significance. When asked to recall significant examples, human experts are often found to apply unequal weighting across time (Epstein and Segal 2000). Some experts prioritize the more recent over more distant past. Others recall policy outputs that are significant from a longer-term perspective. With our web-based approach, we are able to minimize such recall bias by performing date-restricted searches and querying only the content posted within a specific time window. This allows us to focus on contemporaneous evaluations that assess significance of a policy output around the time of its enactment.

Learning from Positive and Unlabeled Examples

Our starting point is thus a set P of policy outputs which have been mentioned in the web content posted by online expert raters and which we consider to be significant in the sense discussed above. The second step in our approach is to find other policy outputs which are sufficiently similar to our significant examples, starting from the seed set of positives, P.

Multiple classification methods exist for learning from positive and unlabeled data, including two-step methods, biased two-class classifiers, statistical queries, and one-class classifiers (Zhang and Zuo 2008). In this work, we introduce an established two-step Rocchio-SVM method proposed by Li and Liu (2003) and Liu (2011) which is widely used in web data mining.[5] We extend and adjust this method for our purposes by (i) incorporating both categorical and textual features, (ii) running multiple models with varying train-validation splits, and (iii) constructing bootstrap confidence intervals for model predictions (see Appendix A).

Our algorithm has three overall steps (see Table 1). We start with a set of policy outputs $L = P \cup U$ where P is the set of positives and U is the unlabeled set. The first step is the detection of reliable negatives $RN \subset U$, i.e. policy outputs that can reliably be classified as not significant. Two sets of reliable negatives RN_T and RN_C are found using textual

[3]We formulate such assumptions for the purposes of our application in Appendix C.
[4]We come back to this point when we discuss possible extensions in the conclusion.
[5]Like all other PU learning methods, this method makes important assumptions - see Appendix A.

and categorical features respectively. Having found RN_T and RN_C, we propose to obtain the final set of reliable negatives as $RN = RN_T \cap RN_C$, i.e. the intersection of the reliable negatives obtained from using the Rocchio method on the textual and categorical features.

Table 1: The PU Learning and Prediction Algorithm

1. Find the reliable negatives $RN \subset U$, using the adapted Rocchio method:
   (a) Find reliable negatives RN_T using textual features:
      i. Calculate the positive centroid $\bar{P}$, and the cosine distance of every output to $\bar{P}$.
      ii. Find potential negatives PN_T as all outputs which are farther than ω from $\bar{P}$, where ω is the farthest distance of any output *in* P to $\bar{P}$.
      iii. Find adjusted positive and negative centroids: $\mathbf{c}_P = \alpha\bar{P} - \beta P\bar{N}_T$, $\mathbf{c}_N = \alpha P\bar{N}_T - \beta\bar{P}$.
      iv. Find the set of reliable negatives RN_T as all outputs which are closer to $\mathbf{c}_N$ than $\mathbf{c}_P$.
   (b) Find reliable negatives RN_C using the categorical features:
      i. Proceed as under (i)–(iv) above, except that: use the Manhattan distance instead of the cosine distance; use the median of the features and set $\mathbf{c}_P = \text{med}(P)$ and $\mathbf{c}_N = \text{med}(PN_C)$.
   (c) Find reliable negatives as $RN = RN_T \cap RN_C$, the intersection of the reliable negatives obtained from (a) and (b) above.
2. Run an iterative SVM classifier using a set of positives P, reliable negatives RN and the remaining set of unlabeled items $Q = U - RN$:
   (a) Split the positives into train positives P_{train} and a validation set P_{validate}.
   (b) Train SVM on the training data P_{train} and RN to classify the unlabeled outputs in Q. Let the predicted negatives be denoted by W.
   (c) If $W = \emptyset$, then proceed to step (e). Otherwise, add W to the reliable negatives RN and drop them from the unlabeled set Q.
   (d) Iterate the process from (b) and (c) above, until (i) no new negatives are found or (ii) prediction accuracy on the validation set P_{validate} falls below 0.85.
   (e) Save the model with predictions and stop.
3. Predict the significance of legislation:
   (a) Train 100 models in the previous step with various splits of P_{train} and P_{validate}.
   (b) For each output $\mathbf{x}_i \in L$, obtain a vector of 100 predictions $\mathbf{g}_i$.
   (c) Sample with replacement from each prediction vector $\mathbf{g}_i$ to take 1000 bootstrap samples, and calculate sample means $\mathbf{\Lambda}_i$.
   (d) Find a_i as the 1st percentile of $\mathbf{\Lambda}_i$.
   (e) Classify policy output i as significant if $a_i \geq 0.5$.

In the second step, the algorithm runs a support vector machine (SVM) classifier iteratively using positives P, reliable negatives RN and the remaining set of unlabeled items, $Q = U - RN$ (Liu 2011, pp. 195-7).[6] In each iteration, a new classifier is trained using P and RN, which is then used to predict labels for all outputs in Q. The outputs that are classified as negative, W, are dropped from Q and are added to RN. The iterated SVM method proceeds by augmenting the RN at each iteration, until no items in Q are classified

[6] We apply an IDF transformation to $\mathbf{X}_T$, first to the training data and then to the validation set.

as negative or the prediction accuracy on a validation set drops below 85% (Liu 2011).

The final step is prediction. To ensure that a particular choice of train-validation split for positives does not bias prediction, we extend the Rocchio-SVM method and run 100 models with different train-validation splits and obtain the mean prediction as a proportion of the 100 models that predicted a policy output as significant. We estimate the uncertainty of this mean using the bootstrap method. We consider output i as significant if the lower-bound of a one-sided 99% percentile bootstrap confidence interval is greater or equal to 0.5.

APPLICATION

UK Secondary Laws

We offer an application of our proposed approach by measuring the significance of secondary legislation in the United Kingdom (UK). Most legislative action in the UK occurs through secondary, rather than primary, legislation (Page 2001). Primary laws contain only a broad regulatory framework, while secondary laws provide details that are too complex to include in primary legislation. The principal form of UK secondary law – and the one which we study in this work – is known as the *statutory instrument* (SI). Figure 1 shows the distribution of SIs by year between 2009 and 2016. A large number of SIs regulate matters in the fields of education, social security, health, income tax and pensions[7]. The question we address in this application is simple: which SIs are significant?

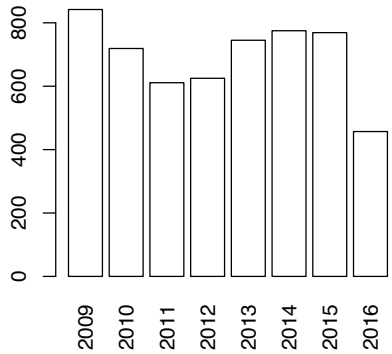


Figure 1: Number of SIs by year

[7]See Appendix B for details on data sources.

Law Firms as Raters

In this application, we source examples of significant laws from the web pages of law firms. Besides providing a means to introduce attorneys and services, websites offer an attractive platform for law firms to demonstrate expertise within their practice areas. Such 'content marketing' takes the form of client alerts, practice-group newsletters and industry-specific blogs, targeted at current and prospective clients. Regulatory updates drawing attention to major changes in legislation are a key part of these marketing activities (Greentarget 2017).

To understand why law firms may determine some laws significant, we review the literature on web content marketing for the legal profession. While it is not possible to identify why law firms write about legislation in all cases, we find a consensus that spans general observations on the current state-of-the-art in legal content marketing, recommendations from professional legal bodies and observations from those working in law firms. Our review indicates that lawyers post content online mainly about laws that *change the regulatory status quo by a large margin* (see Appendix C). Our set P can thus be expected to contain UK statutory instruments that are significant in this sense.[8]

To ensure we select law firms that have the resources and expertise to engage in professional content marketing, we focus on the web pages of top-ranked, leading UK law firms only. By focusing on leading law firms, we also have the benefit of choosing well-informed, highly specialist experts, as top firms retain the best legal professionals with a broad and deep understanding of the regulatory environment. In this study, we select 288 top-ranked, leading UK-based law firms drawing on two of the largest reputable rankings of the legal profession: *Chambers & Partners* and *The Legal 500* (Appendix D).

Web Search

The 288 firms in our selection are associated with 317 websites. Each law firm is associated with at least one website. Based on this set of websites, we construct a custom search engine using the Google Custom Search facility[9]. As our goal is to retrieve content that is relevant to a specific law, we place a larger penalty on retrieving false positives than false

[8]This expectation holds in our data - see next section.

[9]See https://cloud.google.com and https://developers.google.com/custom-search.

negatives. Consequently, we use the full quoted title of a law as our query. We control the temporal dimension of law significance by performing date-restricted searches with the date on which a web page was created as an anchor. Our search queries the web content that law firms posted in the six-month period preceding the law's publication and in the six-month time period following its publication.

In August 2017, we searched for the quoted title of each of 9,838 SIs adopted between 1 January 2009 and 31 December 2016. Our search results are an SI-firm matrix, containing hit counts for each SI in each firm's website. As we are not interested in the exact hit counts within each firm's website, we binarize this matrix. To construct output set P, one must determine a threshold of web citations across firms above which one will consider a law significant (i.e. choose P_n – see previous section). In this application, we set $n = 2$, thus choosing all SIs for which we find relevant content on at least two firms' websites.[10]

Our output set P_2 has 271 laws. The set has high face validity as a sample of laws that can be considered as significant in the eyes of law firms, i.e. laws that change the regulatory status quo by a large margin. An inspection of the set P_2 reveals that, compared to the set U, the laws in P_2, on average, contain longer texts, more often introduce new provisions and have wider regional application. This pattern holds regardless of whether the set P_2 is obtained from the pages of firms that are small or large, regional or London-based, specialized or full-service (see Appendix E).

Model Training

To train the PU learning model, we first select features focusing on information which can be plausibly related to output significance. We use both real-valued textual and categorical features.[11] The former are obtained from the text of Explanatory Notes which are attached to each UK statutory instrument. We use bigrams as tokens, as they are more informative than unigrams[12]. Our categorical features include: the *major topic*, denoting the subject matter into which an SI falls; the *department* (ministry) responsible for adopting it; a battery of features that capture the different categories of SIs; the *location* capturing an

[10]We re-run the analysis using two other thresholds: $n = 1$ and $n = 3$, and also by varying the number of law firm pages. See Appendix G.

[11]See Appendix F for details on features, data preparation and models with partial feature sets.

[12]Key phrases denoting significance, such as 'no impact', can be captured more effectively by bigrams.

SI's geographical coverage; the *SI length*; the feature *laid-before-parliament* which denotes whether an SI was submitted for legislative scrutiny; and finally two features – *memorandum* and *impact assessment* – capturing documents that can accompany a statutory instrument.[13]

We train our model using the PU learning algorithm from Table 1 using the positive set P_2 which has 271 laws and the set of all unlabeled laws adopted between 2009 and 2016.[14] In the first step, our algorithm finds reliable negatives. Our RN_T contains $5,830$ laws, RN_C contains $2,272$ laws, and the intersection RN has $2,089$ laws. The unlabeled set Q contains $7,194$ laws. In the second and third step, we train one hundred iterated SVM models and for each law we obtain one hundred predictions of the label. We then take 1000 bootstrap samples of the set of predictions and find the 1st percentile of the distribution of the bootstrap means to determine significance. We consider any law for which this lower bound exceeds or is equal to 0.5 as significant.

In supervised learning with labeled positives and negatives, we can use standard methods of determining model fit such as accuracy or AUC. As we only have true positives in PU learning, we estimate model fit by the proportion of positives correctly predicted. We correctly predict 251 out of 271 positives, or with 92.6% accuracy, which shows our model does not underfit.[15] To understand how true positives are distributed according to confidence of prediction, we examine the density of the lower bound of the 99% confidence interval (a_i) that we use to determine significance. Figure 2 shows the density of the lower bounds a, demonstrating that our model is confident to a high degree for the large majority of positives.

We can consider feature importances to understand the key predictors of significance.[16] The top predictive features, both contributory and inhibiting, are shown in Table 2, with * denoting an interaction. The list is dominated by non-textual features (see Appendix H for a discussion of text feature importances). The length of the SI is a major predictor, appearing in four out of the ten top contributing features, as itself and in interaction with other features. Whether an SI is a regulation, has an impact assessment, or has a memorandum, are also

[13]For categorical features, we use second-degree polynomial expansion to consider all possible feature interactions, except self-interactions.

[14]We implement the algorithm in Python 3.5.

[15]This is the accuracy obtained from training data. For test accuracy, see next section.

[16]We use the coefficients from the model to determine feature ranking (Chang and Lin 2008). As we have 100 models, we use mean rank to determine a feature's relative contribution to importance.

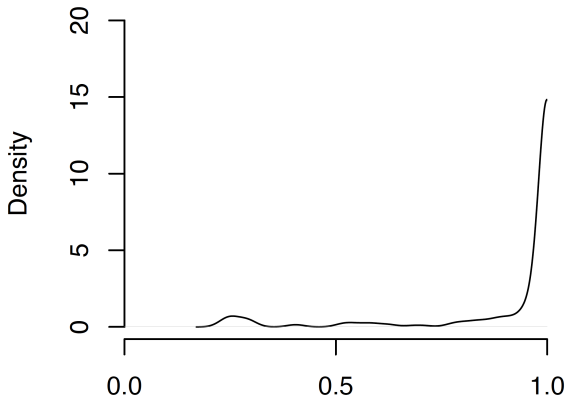


Figure 2: Density of the lower bounds a.

good predictors of significance. Major topics such as labor and domestic-commerce are good predictors. SIs with one or more of the above contributing factors present are thus more likely to be classified as significant.

EVALUATION

We evaluate our method in three ways, starting with two tests on unseen data of all UK statutory instruments adopted between 1 January and 31 July 2017, beyond the 2009-2016 period from which we obtained our training data. The third test probes the construct validity of our significance estimates through an analysis of the UK legislative cycle in 2005-2015.

Forecasting Citations

Our first evaluation offers a simple test. Is our PU learning model able to predict which SIs law firms will consider sufficiently important to mention on their websites? An answer in the positive would indicate that our model performs well in its domain and effectively forecasts citations of laws within web page content. More specifically, it would show that our model has good recall, i.e. it correctly retrieves a high fraction of true positives.

To check this, we performed a new Google search in February 2018 to obtain web citations of SIs published in the first half of 2017. We found that out of the total of 409 SIs made in that period, 26 are cited on the websites of at least two law firms. Out of the 26, our model correctly predicts 22 positives (84.6%), misclassifying four SIs (Appendix J). While we obtain

Table 2: Feature importances

contributing feature	mean rank
regulations * memorandum	5.790
labor * memorandum	7.420
regulations * si-length	16.350
impact-assessment * si-length	18.390
memorandum * si-length	21.060
impact-assessment	22.430
si-length	38.070
department/business * impact-assessment	39.425
domestic-commerce * regulations	45.590
location/uk * memorandum	45.890
inhibiting feature	**mean rank**
amendment	3.960
department/justice * amendment	14.650
transitional * memorandum	16.345
regulations * amendment	16.820
government-operations * amendment	18.615
macroeconomics * location/one-region	31.595
department/communities-local-government * macroeconomics	36.615
commencement * si-length	36.630
department/justice * location/uk	72.220
rules * amendment	94.230

a high true positive rate, we do predict more positives than the actual number of positives, as our model classifies 155 (37.9%) of the 409 SIs as significant.[17] This is not necessarily a problem, as the new set P_2 obtained for 2017 is an incomplete set of significant outputs. To check whether we achieved a high recall just by chance, we calculate the probability of this happening and obtain a very low probability of 2.98×10^{-6} (see Appendix K).

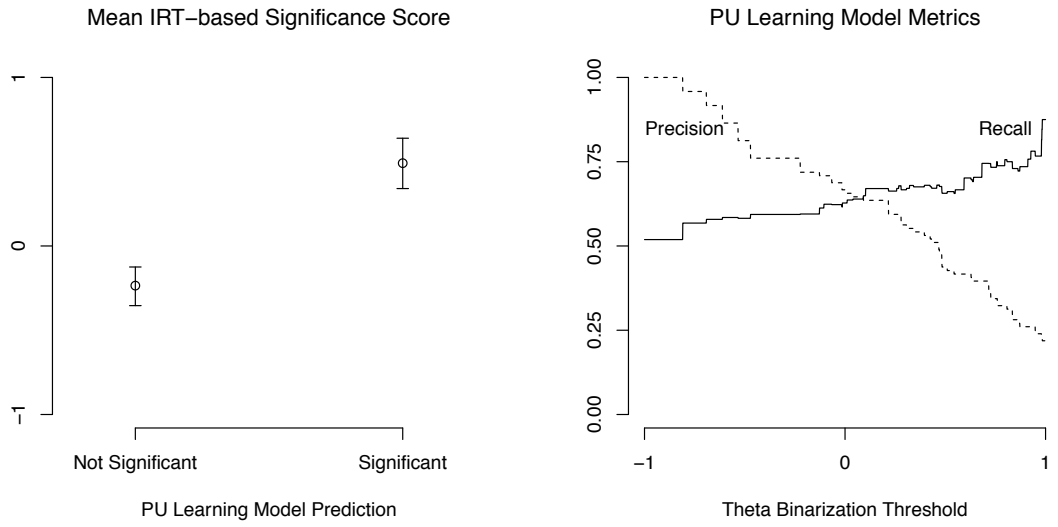
Comparison to Human Experts

For our second evaluation, we compare our classification results for 2017 with estimates derived from human experts. We obtain expert ratings for a sample of 217 of the 409 SIs from three legal professionals (Appendix I). Using the Qualtrics survey platform, we requested that each expert reads the full text of each law, before prompting them as follows:

[17] One prominent study (Page 2001) estimates roughly that about 30% of UK SIs are *politically* important.

'Imagine that you are a lawyer working for a major UK law firm and you are tasked with selecting important pieces of legislation (Statutory Instruments) to write about for the firm's clients. On a scale from 1 to 6, how important do you think this SI is to write about?'

Based on our expert rankings, we estimated a graded response IRT model to derive the theta values of latent significance (Clinton and Lapinski 2006).[18] For validation purposes, we consider these theta values as the true (continuous) estimates of SI significance. We first compare the mean theta values (with 95% bootstrap CIs) for SIs that our model predicts as not significant and significant. Figure 3(a) reports the results. As this figure shows, there is a strong correspondence between human experts and our classification. The mean theta-based significance score for the SIs predicted as not significant is -0.35 and that for the SIs predicted as significant stands at 0.44 (where the overall mean theta is 0.00).



(a) **(b)**

Figure 3: Comparison with Human Experts

In the second step, we can consider the precision and recall of our PU learning model. To do so, we must binarize theta values as IRT-based significance is measured on a continuous scale. Rather than choosing an arbitrary binarization threshold, we examine how precision and recall change as one varies the threshold from -1 to 1 on the theta scale. Figure 3(b) presents the results. This analysis reveals that our model has best accuracy when one bina-

[18] We discuss inter-rater reliability and results with alternative score aggregation methods in Appendix I.

rizes on theta-based significance values between 0.103 and 0.216 range (Accuracy 0.70, Recall 0.67, Precision 0.64). In an error analysis we find that our model tends to misclassify more nuanced SIs, an issue it shares with other automated classification methods (see Appendix I for a discussion of the types of errors that our model makes compared to human experts).

Significant SIs and UK Legislative Cycle

The previous two evaluations show that our model fairly effectively replicates citations of legislation by leading UK law firms, both when tested with web data and with human expert ratings. For our final evaluation, we investigate whether our predicted labels accord with the general notion of significance as a major departure from the regulatory status quo. To probe this construct validity of our significance estimates, we examine how the share of the SIs we classify as significant varies over the annual legislative cycle.

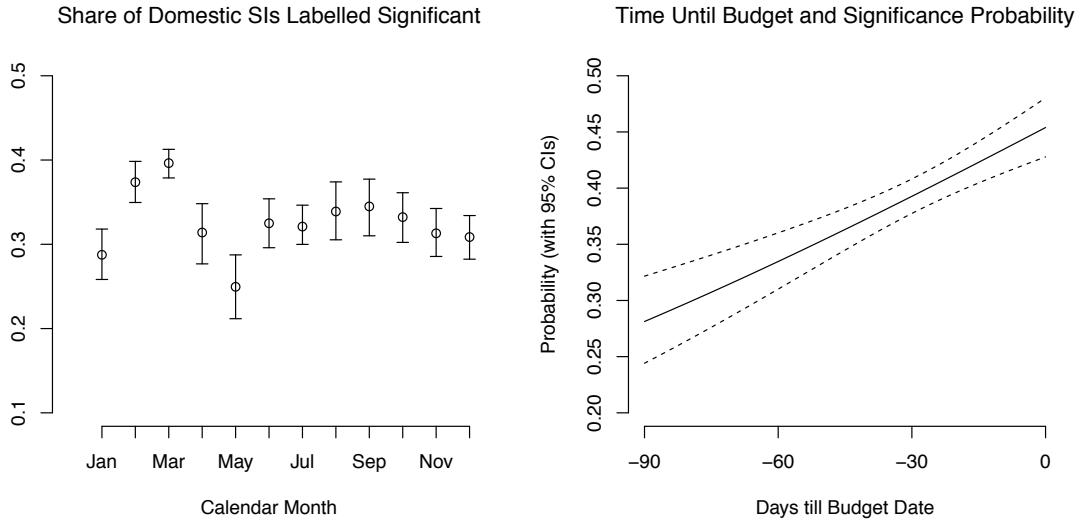
In the UK, a high number of statutory instruments are adopted over a few weeks before the budget is announced in late March of each year.[19] The constitution permits large parts of the budget to come into force almost immediately once announced, and hence secondary legislation must be ready to give substantive effect to many finance bill clauses (Seely 2020). The budget is a major policy event and, if our classification has good construct validity, we expect that a relatively high share of statutory instruments passed in the weeks prior to the budget will be classified as significant. In fact, given the focal role that the budget plays in a domestic policy cycle, we would expect to find few other periods in any given year when the proportion of significant SIs matches that during the weeks leading to the budget.

This is exactly what we find. We classify all statutory instruments adopted under the Blair/Brown government (2005–2010) and the Cameron-Clegg coalition (2010–2015), limiting our analysis to domestic SIs which are influenced by the UK budgetary cycle.[20] Figure 4(a) shows how the monthly share of domestic SIs that our model predicts as significant varies over the calendar year. This analysis reveals a marked surge in the production of significant SIs in February and March, the two months before the budget. In a separate analysis, we use the number of days until the budget date (in the 0–90 range) as a predictor

[19]The budgetary cycle was reformed in 2017.

[20]See Appendix L for results with alternative samples and specifications. We note in passing that the results for SIs that are not subject to the UK budget cycle support our argument.

in a generalized linear model with binomial distribution and a logit link, in which the dependent variable is the significance label obtained from our model. Figure 4(b) reports the results. As this figure demonstrates, the significance of SIs, as captured by our classification, increases strongly as one moves closer to the budget date. We believe this shows that our estimates have good construct validity.



(a) **(b)**

Figure 4: Statutory Instruments and UK Budget Cycle

CONCLUSION

In this paper, we proposed a new method for measuring the significance of policy outputs, based on a semi-supervised machine learning approach. We show that with this approach we can obtain estimates that fairly accurately accord with future web content and hand-coded expert ratings and also have more general construct validity. There is a trade-off to our method. Our web-based technique is efficient and automated but it may struggle with some of the more nuanced cases, thereby lowering prediction accuracy. As ours is a proof-of-concept work, we leave the task of further improving model performance for future work. Our approach allows for interesting extensions. By performing searches on various subsets of web resources, researchers can train models using examples of significant outputs from expert raters who use distinct concepts of significance. We showcased the method using

'seed' sets from professional services firms that focus chiefly on outputs shifting the status quo by a large margin. Future research may train models on examples obtained from raters who use other concepts of significance. Outputs that are consistently classified as significant across such models may be assumed to have general significance.

References

Angelova, Mariyana, et al. 2018. "Veto player theory and reform making in Western Europe". *European Journal of Political Research* 57 (2): 282–307.

Becher, Michael. 2010. "Constraining Ministerial Power: The Impact of Veto Players on Labor Market Reforms in Industrial Democracies, 1973-2000". *Comparative Political Studies* 43 (1): 33–60.

Chang, Yin-Wen, and Chih-Jen Lin. 2008. "Feature ranking using linear SVM". In *Proceedings of the Workshop on Causation and Prediction Challenge*, 53–64.

Clinton, Joshua D, and John S Lapinski. 2006. "Measuring legislative accomplishment, 1877–1994". *American Journal of Political Science* 50 (1): 232–249.

Conley, Richard S., and Marija A. Bekafigo. 2010. "No Irish Need Apply? Veto Players and Legislative Productivity in the Republic of Ireland,1949-2000". *Comparative Political Studies* 43 (1): 91–118.

Döring, Herbert. 2001. "Parliamentary Agenda Control and Legislative Outcomes in Western Europe". *Legislative Studies Quarterly* 26 (1): 145–165.

Epstein, Lee, and Jeffrey A. Segal. 2000. "Measuring Issue Salience". *American Journal of Political Science* 44 (1): 66–83.

Greentarget. 2017. *State of Digital & Content Marketing Survey*. Greentarget-Zeughauser Group, http://greentarget.com/clients/new-media-engagement-survey.

Howell, William, et al. 2000. "Divided Government and the Legislative Productivity of Congress, 1945-94". *Legislative Studies Quarterly* 25 (2): 285–312.

Huber, John D, and Charles R Shipan. 2002. *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. CUP.

Junge, Dirk, Thomas Koenig, and Bernd Luig. 2015. "Legislative Gridlock and Bureaucratic Politics in the European Union". *British Journal of Political Science* 45:777–797.

Li, Xiaoli, and Bing Liu. 2003. "Learning to classify texts using positive and unlabeled data". In *Proceedings of the 18th International Conference on Artificial intelligence*, 587–592.

Liu, Bing. 2011. *Web Data Mining*. Springer.

Mayhew, David R. 1991. *Divided We Govern*. Yale University Press.

Page, Edward C. 2001. *Governing by Numbers*. Hart Publishing.

Schultz, Mike, et al. 2013. *Professional Services Marketing*. John Wiley & Sons.

Seely, Antony. 2020. *The Budget and Annual Finance Bill*. Briefing Paper 813. HC Library.

Tsebelis, George. 1999. "Veto players and law production in parliamentary democracies: An empirical analysis". *American Political Science Review* 93 (03): 591–608.

Warber, Adam L., Yu Ouyang, and Richard W. Waterman. 2018. "Landmark Executive Orders". *Presidential Studies Quarterly* 48 (1): 110–126.

Zhang, Bangzuo, and Wanli Zuo. 2008. "Learning from Positive and Unlabeled Examples: A Survey". In *2008 International Symposiums on Information Processing*.