Measuring the Significance of Policy Outputs with Positive Unlabeled Learning

Online Appendix

Radoslaw Zubek, Abhishek Dasgupta, David Doyle

31 August 2020

Contents

A Positive-Unlabeled Learning

A.1 Method Selection

Positive unlabeled (PU) learning refers to a class of semi-supervised learning algorithms that learn from data which has only one class, that of positives. As mentioned before, there are various approaches to PU learning: two-step methods, biased two-class classifiers, statistical queries, and one-class classifiers (Zhang and Zuo 2008).

Two-step methods (Li and Liu 2003; Liu 2011; Fung et al. 2005) such as the one used in this paper, as the name implies, comprise two steps: the first is the extraction of reliable negatives from the unlabeled set, with the second being an iterative classification method which trains on the positives and extracted negatives. Variations in the two-step method come from different methods of extraction of reliable negatives, such as by using nearest-neighbour classifiers instead of Rocchio; while the variation in the second step can come from selecting different classifiers or different stopping criteria for the iterative classifier.

Biased classifiers (Liu et al. 2003; Lee and Liu 2003) treat the problem as a supervised classification problem but with the positives and unlabeled weighted differently. These methods generally assume that the unlabeled set contains mostly negatives (true for large sample sizes). This assumption can be encoded by varying constraints in the optimisation method and varying the relative error costs for false positives (lower cost) and false negatives (higher cost), or by weighting the unlabeled sample.

Statistical queries (SQ) is an approach which models statistical properties of datasets. Work in this area (Denis, Gilleron, and Letouzey 2005) generalises SQ-like algorithms to a learning model for positive queries.

One-class classifiers (Schölkopf et al. 2001; Manevitz and Yousef 2001) also use binary classifiers. They work by constructing a function which separates out the (usually relatively small) region of feature space that has the positive class. This gives a separation in feature space using only one class. This is accomplished by first transforming the feature space. Then a specific formulation of a standard classifier is used which treats the origin as the only member of the second class to give the separation in feature space.

In this work, we utilize an established two-step Rocchio-SVM method proposed by Li and

Liu (2003) and Liu (2011). This method is a preferred choice for our purposes as it has been widely used in web data mining. We adjust this method in three important ways. First, as we have both categorical data and text data, we extend the method to handle both types of features. Second, we run multiple models with varying train-validation splits to ensure that a particular choice of train-validation split does not bias prediction. Third, we use a model ensemble and obtain bootstrap confidence intervals for our predictions.

A.2 Assumptions

Like other two-step methods, the Rocchio-SVM method makes important distributional assumptions about positives that are reasonable to assume in most cases, including our study. We assume we have a set of outputs $L = P \cup U$ where P and U are disjoint, with each item i characterized by some feature vector $\mathbf{x}_i \in \mathcal{X}$ where $\mathcal{X}$ is the global feature space. Each output has a (hidden) true label $y_i = 1$ if it is positive (significant) and $y_i = -1$ if it is negative (not significant), and the output $(\mathbf{x}_i, y_i)$ is sampled from a joint probability distribution $D_{\mathbf{x},y}$. Our objective is to learn a function f which maps the features $\mathbf{x} \in \mathcal{X}$ to the label $y \in \{-1, 1\}$ such that the classification error is minimised. The PU learning approach further assumes that the set of positives P is a sample from a conditional distribution $D_{\mathbf{x}|y=1}$, which is the global feature distribution of all the positives. In other words, we assume our positives P and the set of unseen positives in U to be from the same distribution. This allows us to infer other positives from the unlabeled set U.[21]

A.3 Algorithm

We start from a $N \times M$ feature matrix $\mathbf{X} = (\mathbf{X}_T, \mathbf{X}_C)$ representing N outputs with M features, where $\mathbf{X}_T$ is the $N \times M_T$ term frequency (TF) matrix of textual features and $\mathbf{X}_C$ is the $N \times M_C$ categorical feature matrix. Our algorithm for modeling significance of policy outputs has three overall steps: (i) finding reliable negatives, (ii) using an iterative classifier, and (iii) prediction (see Table 1).

[21] This last assumption implies that the nature of positives does not change over the time period of observation which is reasonable to assume in most cases, unless the observation period is very long.

Detection of reliable negatives

The first step in our algorithm is the detection of reliable negatives $RN \subset U$, i.e. policy outputs that can reliably be classified as not significant. We need the set of reliable negatives to be able to train a classifier in the next step. The standard Rocchio-SVM method uses only real-valued (textual) features in this first step. As we have both categorical data and text data, we adjust the standard method to handle both types of data. We find two sets of reliable negatives RN_T and RN_C using textual and categorical features respectively. To find RN_T, we use the Cosine-Rocchio method (Liu 2011), a variant of an early method of text classification (Rocchio 1971). First, we convert the rows $\mathbf{d} \in \mathbf{X}_T$ to a normalised TF-IDF scheme. Let the centroid of a set of rows $S = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K\}$ with $K \leq N$ be defined as:

$$\bar{S} := \frac{1}{K} \sum_{i=1}^{K} \mathbf{d}_i \tag{2}$$

We calculate the cosine distance of all rows in $\mathbf{X}_T$ to the positive centroid $\bar{P}$, where $\bar{P}$ is the centroid of the set of rows in $\mathbf{X}_T$ that correspond to our set of positives P. For two normalised vectors $\mathbf{x}$ and $\mathbf{y}$, the cosine distance $\cos(\mathbf{x}, \mathbf{y})$ is defined as:

$$\cos(\mathbf{x}, \mathbf{y}) = 1 - \mathbf{x} \cdot \mathbf{y} \tag{3}$$

These distances are used to find all outputs in U that lie beyond a certain threshold distance ω from $\bar{P}$, where ω is the distance of the farthest positive from the positive centroid. This set of outputs, denoted as PN_T, is our set of *potential negatives*. To set the relative impact of positives and negatives, we obtain the adjusted positive centroid $\mathbf{c}_P = \alpha\bar{P} - \beta P\bar{N}_T$ and the adjusted negative centroid $\mathbf{c}_N = \alpha P\bar{N}_T - \beta\bar{P}$, where $P\bar{N}_T$ is the centroid of the potential negatives. As recommended in (Liu 2011), we set $\alpha = 16$ and $\beta = 4$. Finally, we obtain the set of *reliable negatives*, RN_T, as those outputs which are closer to the $\mathbf{c}_N$ than $\mathbf{c}_P$.

To find reliable negatives RN_C using the categorical features, we again adjust the Rocchio method employed in Liu (2011). Most importantly, we use the median of rows in $\mathbf{X}_C$ instead of the centroid, since the mean would give us a feature vector that is no longer categorical.

For a set of rows $S = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K\}$, let the median be defined as

$$\text{med}(S) = (\text{med}(\mathbf{d}_{1,1}, \ldots, \mathbf{d}_{K,1}), \ldots, \text{med}(\mathbf{d}_{1,M_C}, \ldots, \mathbf{d}_{K,M_C})) \tag{4}$$

Also, for categorical features, the use of cosine distance is not appropriate as the vector is binary. We use the Manhattan distance in the Rocchio instead of cosine distance, where the Manhattan distance (also known as L_1) between two vectors $\mathbf{x}, \mathbf{y}$ is defined as:

$$L_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M_C} |\mathbf{x}_i - \mathbf{y}_i| \tag{5}$$

We find the potential negatives set PN_C using distances to the positive median and the threshold ω. We set $\alpha = 1, \beta = 0$, and hence obtain $\mathbf{c}_N = \text{med}(PN_C)$ and $\mathbf{c}_P = \text{med}(P)$. Then RN_C are all outputs which are closer to $\mathbf{c}_N$ than $\mathbf{c}_P$.

Having found RN_T and RN_C, we can proceed 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.

Using an iterative SVM classifier

In the second step of the algorithm, we use the support vector machine (SVM) classifier. SVM is a widely used binary classifier (Hastie, Tibshirani, and Friedman 2009), not least because it is not sensitive to outliers. SVM classifies by constructing a hyperplane (plane in high dimensions) that separates the positive and negative classes. Formally, given training data $\langle(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_M, y_M)\rangle$, with $y_i = 1$ for positive instances and $y_i = -1$ for negative instances, SVM learns a function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ where b is a bias term. Sometimes there may not be a separating hyperplane between the positive and negative classes; we must then allow for some misclassified instances on either side of the hyperplane. The degree of allowed misclassification is given by the regularization parameter C, with ξ denoting

misclassification error. The SVM formulation is then an optimization problem:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}\xi_i \\ &\text{with constraint} && y_i \times (\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1, 2, \ldots, M \end{aligned} \tag{6}$$

We run SVM iteratively using positives P, reliable negatives RN and the remaining set of unlabeled items, $Q = U - RN$ (Liu 2011, pp. 195-7).[22] In each iteration, a new classifier is trained using P and RN, which is then used to predict labels for all documents in Q. The documents that are classified as negative, W, are dropped from Q and are added to RN. If W is empty, then we save the model with predictions and stop. Otherwise, a new SVM classifier is constructed using positives P and the new augmented RN. The iterated SVM method proceeds by augmenting the RN at each iteration, until no items in Q are classified as negative.

At each SVM iteration, we determine the best regularization parameter C by doing a 5-fold cross validation with a grid search over a range of values.[23] To perform k-fold cross-validation, k splits of the training data are obtained, and $k-1$ splits are used for training and the remaining split are used as a validation set (Hastie, Tibshirani, and Friedman 2009, section 7.10.1). The train-validation splits are varied so that each split gets to be tested once, giving us k training sets. For a single value of C, we can estimate model accuracy by averaging the validation accuracy across these models. Then a grid search is performed for all the values of C to determine the best value of C by model accuracy. This C is then used for training.

There is some debate whether the classifier at the convergence is the best model. As recommended in (Liu 2011), we use the prediction accuracy on a validation set of positives to inform our choice of the best model. We thus keep a held out set of positives (20%) as a validation set and construct the SVM classifier using the rest (80%) at each iteration. This validation set is distinct from the validation set used for k-fold cross validation. If the accuracy falls below a certain threshold (85%) then we stop the iteration. Our result is thus

[22] Before we run an SVM, we apply an IDF transformation to $\mathbf{X}_T$, first to the training data and then to the validation set.

[23] We use a C grid of $\{10^{-2}, 10^{-1}, \ldots, 10^{10}\}$.

the model obtained when: (i) no new negatives are obtained or (ii) the prediction accuracy on the validation set drops below our pre-determined threshold.

Prediction

Predicting the class in a SVM is the computation of $f(\mathbf{x})$. If $f(\mathbf{x}) > 0$ then a policy output is positive (has significance), otherwise it is negative (not significant).

It is possible that a particular choice of train-validation split for positives may bias the prediction. As this is not considered in the standard Rocchio-SVM algorithm, we adjust the method by running 100 models with different train-validation splits. Thus, for each of our policy outputs in $L = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, we get 100 predictions giving us a $100 \times N$ prediction matrix $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N)$ where $\mathbf{g}_i$ are the columns of $\mathbf{G}$. Here the elements of the matrix are 1 or 0 corresponding to the SVM classifications of 1 and -1 respectively. The predictions of the 100 models can be considered as a sample of predictions from a population of all models based on a 80-20% train validation split.

We obtain the mean prediction as the proportion of the 100 models that predicted a policy output as significant. In a further departure from the standard Rocchio-SVM algorithm, we estimate uncertainty of this mean using the bootstrap method. We draw 1000 bootstrap samples from the prediction vector $\mathbf{g}_i$ for each policy output and calculate the bootstrap means $\mathbf{\Lambda}_i$ for an output i. This gives us a $1000 \times N$ matrix $\mathbf{\Lambda} = (\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \ldots, \mathbf{\Lambda}_N)$, where $\mathbf{\Lambda}_i$ are the 1000 bootstrap means for each output. Then we calculate the 1st percentile, a_i, of $\mathbf{\Lambda}_i$ which gives us the lower bound of a one-sided 99% confidence interval. We consider policy output i as significant when $a_i \geq 0.5$.

B UK Statutory Instruments

We obtain our dataset of UK Statutory Instruments from the UK Legislation site (www.legislation.gov.uk). We used the bulk download facility available at http://leggovuk.s3-website-eu-west-1.amazonaws.com. Legislation is published as open data in XML format using the Crown Legislation Markup Language (CLML). We perform parsing and cleaning of the XML data to obtain our categorical features and department, and use the text from

the explanatory notes as the basis for constructing text features. Each SI has a subject which is assigned by the ministry responsible. This subject is usually picked from a pre-existing index of SI subject headings (HMSO 2006, section 2.3.6). We downloaded SIs from 2005–2017, with 2009–2016 being our training set, and data from the first half of 2017 our test set. We obtain 9,838 SIs from the 2009–2016 period, and 460 SIs from the first half of 2017. After dropping SIs which are missing some categorical features (act majortopic, department, subject and type), we are left with 9,554 SIs from 2009–2016 and 409 SIs in 2017. These SIs cover a variety of subjects: we show the top 20 most frequent SI subjects in Table B.1.

subject	frequency
education	1014
social security	883
national health service	685
income tax	620
pension	570
road traffic	531
local government	508
criminal law	505
corporation tax	418
children young people	381
highways	356
police	335
environmental protection	317
immigration	315
financial service markets	258
town country planning	227
healthcare associated professions	221
housing	217
defence	209
transport	206

Table B.1: Most frequent SI subjects

C What Do Lawyers Write About Online?

Although it is not possible to systematically identify why solicitors and lawyers write about legislation on their websites in all cases, a review of what information we do have about content marketing for the legal profession strongly indicates a consensus: they are

concerned with new developments in law or regulations, which amount to deviations from the status quo that would have some form of commercial impact on their clients. This consensus spans general observations on the current state of the art in legal content marketing, recommendations on content marketing from professional legal bodies, observations from lawyers and those working in law firms themselves, to experts writing guides on content marketing for the legal profession.

The websites and blogs of law firms, of all sizes, have become an indispensable tool for content marketing for the legal profession. As Smith (2014, 19) notes: "Firms that blog receive 55 per cent more website visitors than those that don't, as well as 67 per cent more business leads". More significantly however, in *Content Marketing and Publishing Strategies for Law Firms*, Furlong and Matthews (2014, 19) "...conservatively peg at 80 per cent the amount of law firm content dedicated to case commentary and legislative updates, with the other 20 per cent dedicated to practical guidelines, implementable checklists, real-life experiences, and suggested pointers to help clients navigate unfamiliar or difficult territory".

This general focus on writing about legislative updates appears to be reinforced by those working in law firms who are developing this content themselves. A marketing executive, Danielle Bishop, at Southampton-based law firm *Lamport Bassitt*, one of our raters, who was extensively quoted by Smith (2014, 67) as part of a case study on successful legal content marketing, noted the following about the use of social media and blogs for law firms in the UK: "There were definitely law firms out there that really inspired me [...] Pinsent Masons and their outlaw.com site was a big influence – it's a really great blog that is properly updated with relevant news. If new legislation's announced, for example, they're letting their clients know about it straight away through social media". In a similar vein, Barnett and Verney (2009, 158) cite US lawyer, Kevin O'Keefe, who runs LexBlog: "The aim is to provide the legal blog's readers with a constantly renewing source of news and insight about that topic [...] Your legal blog does the work, compiling and passing along updates as they occur".

The notion that lawyers are primarily writing about new regulatory developments and deviations from the existing legal status quo for their clients is further strengthened when one reads the extant work that deals with content marketing for the legal profession. Smith (2014, 22) advises that "if a legal matter has cropped up in the media, seize the opportunity

to explore it on your blog - your reader will enjoy a post that approaches a newsworthy topic from a different angle". Similarly, Monk (2008, 212), when discussing law firm newsletters that can subsequently be uploaded to a firm's website, suggests that subjects can be introduced in a "... conversational low key way, such as changes in the law and how they affect your clients". Monk (2008, 214) further recommends that a general newsletter should be "...informative, telling clients about the latest changes in the law and how it affects them, changes in the environment or taxation regimes, etc." Barnett and Verney (2009, 177), when considering how best to optimise online content on legal websites, suggest regular updates that "...are most likely to take the form of client newsletters - produced at regular intervals, usually monthly or quarterly – and briefing notes, which follow an event such as a key judgement or the announcement or implementation of new legislation".

The Law Society, the representative body for solicitors in England and Wales, also recommends a focus on changes to the legal status quo when developing legal content marketing for the first time. In *Smarter Legal Marketing*, published by the Law Society, Brushfield (2018, 101) suggests that legal newsletters or bulletins are "...a useful way of keeping in touch with your contacts and provide a common marketing vehicle for legal updates". Similarly, in a practice management publication for The Law Society, on internet marketing for your law firm, Adam (2002, 269) recommends a news section on your law firm's website, which should include "...summaries of new legislation or announcements about issues that might affect your clients".

Taken as a whole therefore, it seems clear that when it comes to online content marketing in the legal profession, lawyers, those working in law firms who are responsible for these websites, experts in the area of content marketing for law firms, in addition to professional legal bodies, all generally agree that these websites should contain updates on new regulations or legislation. That is, these websites should primarily consider deviations from the legal status quo, which may have some form of monetary or commercial impact on their clients. As such, we consider significant legislation in line with this understanding, as legislation that causes notable changes or alterations to the existing legal status quo.

D Law Firm Selection

Our sources are two reputable rankings of the legal profession: *Chambers & Partners* and *The Legal 500*. Specifically, we select law firms from the section titled 'A to Z of Law Firms' in *Chambers UK*, and from the sections titled 'Solicitors [England and Wales/Scotland/Northern Ireland/Offshore] A to Z' in the *The Legal 500*.

Before choosing firms to include in our set of raters, we first have to ensure that our firms have clearly defined identities. It is common for large legal firms to merge, split, or be renamed. If such a firm transformation occurs during the time period under consideration, then our rater can no longer be considered to have a continual existence throughout the duration, and are considered as separate firms in our analysis. For our purposes, we assume that a firm has been in continual existence unless both its name and website URL have changed at the same time. Once we have established law firm identities, we first search for firms that have been ranked in every annual edition of *The Legal 500* between 2009 and 2016. We then proceed to do the same with the *Chambers UK*.

We start with the set of firms in the *The Legal 500* 2009 list and assign each firm a unique number called the firm ID (FID). The legal directories also maintain a record of the website associated with a firm for that particular year, and we record this information as a set of tuples of (firm, year, website, FID, source), where source is either The Legal 500 or Chambers. We then proceed by year and add tuples to the list, while determining the FID in each case. The FID is determined as follows: if the firm in question has appeared before (determined by having the same name) then the FID already assigned previously is used. If the firm name has changed but the website has remained the same, we do the same, by looking up the FIDs associated with that website. In both these cases, we only assign an FID if a unique existing FID is found. In all other cases, we assign a new FID.

Once we have added all the years from *The Legal 500*, we proceed to do the same with the *Chambers UK* rankings. In this case, we do not start from an empty list but use the list of (firm, year, website, FID) obtained from Legal500 to build our list, while changing source to Chambers. Our method of assigning FIDs ensures that firms which have merged or split get a different ID, and thus acts as an identifier for a firm.

Finally, we keep only those FIDs which have appeared in all years in either *Chambers* or in all years in *The Legal 500* or in both. The set of websites associated with FIDs gives us the list of raters. A single FID can have multiple websites and multiple names associated with it, due to having changed their website address and/or names.

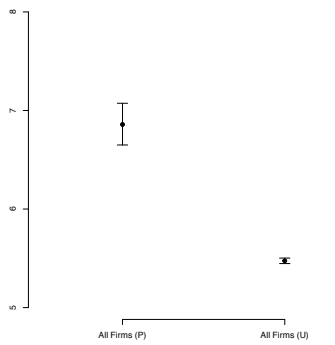
The list of selected law firms can be found in the online supplementary material.

E Positives and Significance

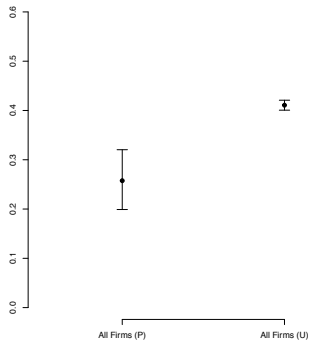
Our review of the secondary literature on web content marketing for the legal profession (see Appendix C) shows that law firms have incentives to post content online mainly about laws that are significant in the sense that they change the regulatory status quo by a large margin. We would thus expect that our P set contains UK statutory instruments that fit this description well. A face validity check indeed shows that the set P_2 contains many such laws, including for example, those regulating taxation (e.g. major changes to the income tax code) and commerce (e.g. new obligations for service providers). Equally importantly, the set appears not to contain laws that introduce relatively minor changes to the status quo.

We further probe the validity of our set P_2 by inspecting whether laws in the P systematically differ from those in the U set. While it is true that the unlabeled set contains many 'hidden' positives, we would expect that the positive set comprises laws that, on average, better fit the description of laws changing the status quo by a large margin. This is exactly what we find. In Figure E.1 (a) - (c), we compare the mean textual lengths and the proportions of amendment SIs and UK-wide SIs in the P_2 and U_2 sets (with 95% bootstrap CIs). This analysis reveals that SIs in the P_2 set are on average longer, are less often amendment SIs, and typically have wider geographical application than the SIs contained in the U set.

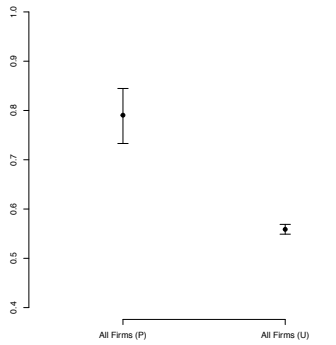
We can also show that these characteristics of the P set hold regardless of the type of law firms from which we source positive examples. We first classify the firms in our selection by size, office location and specialization. A firm is classified as large if it has an above average number of partners, otherwise it is small. A firm is considered London-based if its only office is located in London, regional if all of its offices are outside London, and mixed if its offices are both in and outside London. A firm is full-service if it has an above average number of business practice areas (according to the Legal 500 classification), otherwise it is specialized.



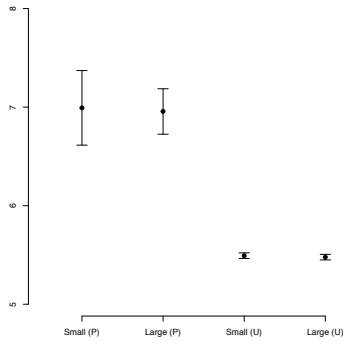
(a) SI Length



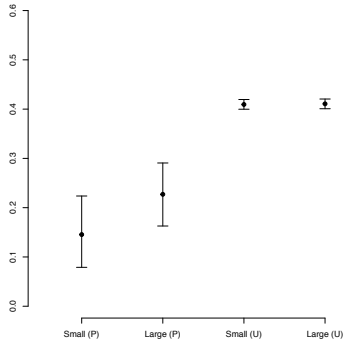
(b) Amendment SIs



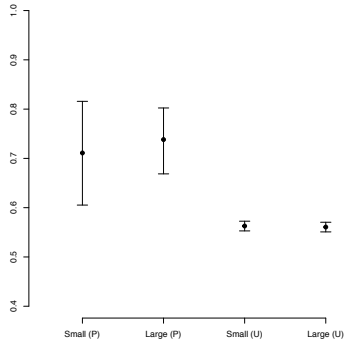
(c) UK-wide SIs



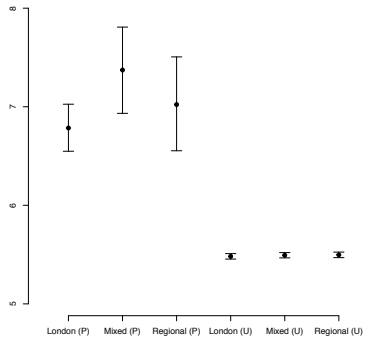
(d) SI Length



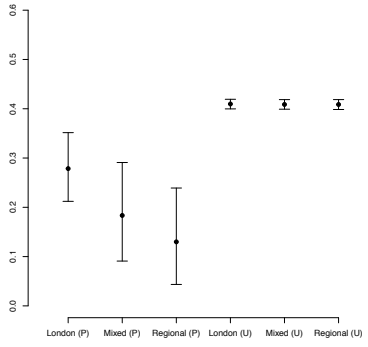
(e) Amendment SIs



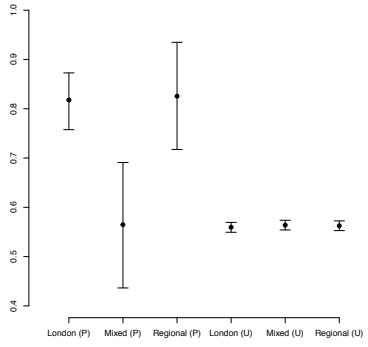
(f) UK-wide SIs



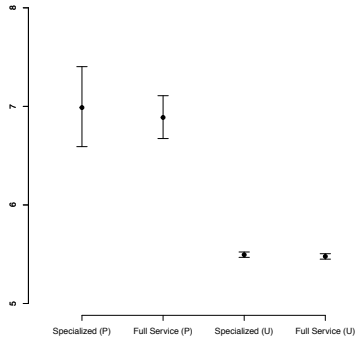
(g) SI Length



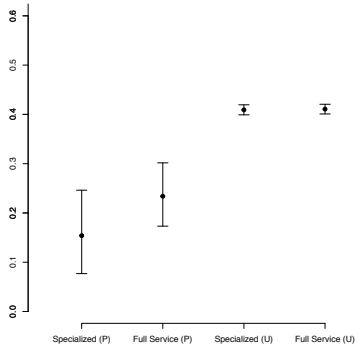
(h) Amendment SIs



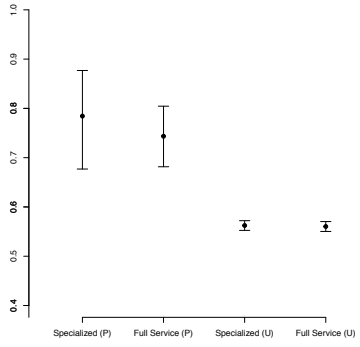
(i) UK-wide SIs



(j) SI Length



(k) Amendment SIs



(l) UK-wide SIs

Figure E.1: Laws in Sets P_2 and U_2 Compared

Figure E.1 (d) - (l) show that the patterns from panels (a) - (c) hold across all firm types. We believe these results provide a strong indication that law firms in our set share a similar concept of significance as a major deviation from the regulatory status quo.

F Features

F.1 Feature Data Sources and Preparation

Our features are of two types: non-textual features which are derived from the attributes of the SI, and textual features from the Explanatory Notes section of the SI. The textual features are the term-frequency (TF) representation of the bag of bigrams from the Explanatory Notes section. This section of an SI is typically a short text describing the legal context of the SI and significant impact. We remove numbers and lemmatize the text. Lemmatization refers to grouping the various forms of a word so they can be analysed as a single term. This helps in reducing the number of unnecessary text features. For the TF representation, we do not restrict the minimum number of SIs a bigram has to appear in, or set the maximum number of bigrams.

The nontextual features are as follows:

- cluster-department: An SI can have multiple departments; we keep all of them. As departments often change their name under new governments with little change in function, we have clustered the departments in our data into coarser categories.[24] The data is encoded as multi-label binary, where each cluster is an independent binary feature.

- act-majortopic: This feature represents the topics associated with an SI is encoded as multi-label binary. The set of topics from an SI is obtained from the topics of the Acts which it references. The topic of an Act is obtained from the Comparative Agendas Project, specifically their categorisation of primary legislation into 22 major topics from 1911 to 2015 (John et al. 2013). As this time period did not cover all the referenced Acts in our database, we developed a system to infer the major topic of an

[24] DepartmentCluster.json in the replication data.

Act (22 possible topics): (i) For each Act, we build an *act subject profile* which is the set of subjects of SIs which reference the Act. (ii) For Acts which do not currently have a topic, we find the set of Acts *with known topic* whose Act profiles exactly match the profile of the Act we wish to find a topic for. (iii) For the exact matched Act profiles, we find the corresponding topic for each of these. This gives us a set of possible topics for the Act. (iv) If the set of possible topics has a unique mode and thus no ambiguity, we assign this topic to the Act. In all other cases, we assign a topic of 0.

- location: The location to which the SI applies as a one-hot encoded feature. We code the location as 0 when the legislation applies UK wide, 1 when it applies to one region, and 2 when it applies to 2 or 3 regions.

- type: SIs are categorised into types such as orders, regulations, rules, byelaws, schemes, among others. Of these types, nearly all SIs are typed as orders, regulations and rules. Thus we keep the type if it is among these three, otherwise we record it as 'other'. We then use one-hot encoding.

- si-length: The number of words in the SI text. This feature tries to capture the effect of length, as significantly short SIs could be expected to have minimal impact. This feature enters our analysis as the total wordcount binarized on the median.

The following are binary features: (a) amendment: whether the SI is an Amendment; (b) commencement: whether the SI is a Commencement Order, a short SI which serves to commence an Act, or a section of it; (c) revocation: whether the SI is an Revocation; (d) transitional: whether the SI is Transitional; (e) transitory: whether the SI is Transitory; (f) laid-before-parliament: whether the SI has been laid before parliament – typically, unimportant SIs are not laid before parliament; (g) memorandum: whether the SI has a memorandum (EM) attached to it – a memorandum is a document attached to some SIs which discusses the context and impact of the SI; (h) impact-assessment: whether the SI has an impact assessment attached – the presence of an impact assessment usually suggests that the SI has some significance.

We drop SIs which have no Acts that have been assigned a major topic.

For categorical features, we also make an adjustment to the categorical feature matrix $\mathbf{X}_C$ before step 1 of the algorithm, i.e. finding reliable negatives, but use the original feature matrix (without the adjustment) for step 2, iterative SVM. The categorical feature matrix comprises both binary (and multi-label binary) and categorical features. We recall that we use the Manhattan distance for finding reliable negatives for the categorical feature matrix. As categorical features are represented as one-hot encoding, the Manhattan distance between two differing SIs which only differed in a categorical feature would be 2. To see why, consider that the two SIs only differ on location. As location has three possible values, its possible one-hot encodings are (1, 0, 0), (0, 1, 0) and (0, 0, 1). Take any two of these and apply the formula for Manhattan distance (Eq. 5 in Appendix A), and we get 2. As binary (and multi-label binary) features would only contribute 1 to the distance this creates an incongruity. We resolve this by multiplying the sub-matrix of the feature matrix that corresponds to categorical features by 1/2. Then, in the case of the location category, we would get the possible feature representations as (0.5, 0, 0), (0, 0.5, 0) and (0, 0, 0.5), giving us a Manhattan distance of 1.

F.2 Analysis with Partial Feature Sets

It is a reasonable question whether we need to train our PU learning model using both categorical and textual features. In this section of Appendix F, we present results for models trained separately on categorical features and textual features. To keep things simple, we undertake this analysis only for the main model with the two-rater threshold ($n = 2$). The main takeaway from this analysis is that, while the model trained only on categorical features produces similar results to that trained on both categorical and textual features, the latter offers a slightly more conservative classification which is closer in the overall proportion of positives to that obtained in Page (2001). In this sense, the full feature-set may be preferable.

Table F.1 presents the key metrics for the training stage for each of the models with partial feature sets and for the full model.[25] The two partial-feature-set models show a fairly good fit on training, although the high true positive rate of the model estimated only

[25]The training set and the number of actual positives differ slightly between the two partial-feature-set models because categorical features are missing for some SIs. See previous section in this Appendix.

with textual features (98.9%) may be indicative of over-fitting. The true positive rate of the model estimated only with categorical features is 86.7% and is below that for the full model. Another metric of interest here is the number of unlabeled SIs which are classified as positive on training. We can see from Table F.1 that the textual-feature-set model is very conservative and classifies relatively few unlabeled SIs as significant on training, while the categorical-feature-set model classifies a relatively large number of unlabeled SIs as positive on training. The results for the full-feature-set model are similar to that for the categorical-feature-set model but are somewhat more conservative. Overall, given these training metrics, we have some preference for the full-feature-set model as it achieves a higher true positive rate, but is slightly more conservative in its classification.

Features	Actual Positives	True Positives	True Positive Rate (Train)	Unlabeled Classified as Positive	Training Set
Textual	277	274	98.9%	830	9,838
Categorical	271	235	86.7%	3,895	9,554
All	271	251	92.6%	3,573	9,554

Table F.1: Model Metrics on Training Data (Two-Rater Threshold, $n = 2$)

Table F.2 shows how the partial-feature-set models perform on web citation forecasting. This is a useful basic evaluation metric for our purposes, as it captures the performance of our models within the primary domain of predicting citations within web content. The true positive rate of the model estimated only with textual features is 66.7% and is much lower than that for the two other models. This is consistent with the evidence of over-fitting in the training stage. The full-feature-set model and the model estimated only with categorical features have the same true positive rate of 84.6%. As expected, the full-feature-set model is slightly more conservative in its classification of positives.

Features	Actual Positives	True Positives	True Positive Rate (Test)	Unlabeled Classified as Positive	Test Set
Textual	27	18	66.7%	46	460
Categorical	26	22	84.6%	142	409
All	26	22	84.6%	133	409

Table F.2: Model Metrics on Test Data (Two-Rater Threshold, $n = 2$)

G Robustness Checks

In this section of the Appendix, we check robustness of our PU model estimation to (i) different thresholds for the minimum number of firms needing to mention a law, and (ii) varying numbers of law firms from which citations are harvested online.

G.1 Analysis with Alternative Thresholds

In the main body of the article, we consider a law to be significant if it is mentioned on the websites of at least two different law firms. Alternatively we could set a different threshold for the minimum number of firms needing to identify a law as significant. In this section of the Appendix, we present results for two other thresholds: (i) at least one firm's website ($n = 1$) and (ii) at least three firms' websites ($n = 3$). The main takeaway point from this robustness check is that our approach performs well with alternative thresholds. It must be noted that because we have less training data with higher values of n, the performance of our PU algorithm declines slightly with larger values of the n parameter.

Table G.1 presents the key metrics for the training stage for each of the three thresholds. All three models demonstrate a good model fit on training. The true positive rate obtained from models estimated with P_1 is 94.6% and that obtained with P_3 is 88.7%. There is thus little evidence of underfitting. Table G.2 presents the metrics for our first simple evaluation test in which we examine the performance of our models in the domain of forecasting citations within web content. All three models demonstrate a true positive rate equal to or above 75%. The rate for models estimated with P_1 is 80.8% and that with P_3 is 75.0%. The model obtained from P_2 has the best true positive rate of 84.6%.

n	Actual Positives	True Positives	True Positive Rate
1	794	751	94.6%
2	271	251	92.6%
3	142	126	88.7%

Table G.1: Model Metrics on Training Data

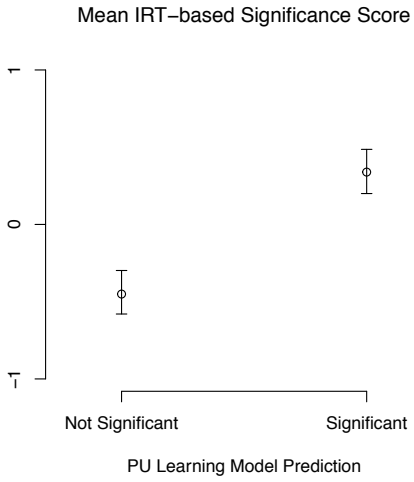
n	Actual Positives	True Positives	True Positive Rate
1	52	42	80.8 %
2	26	22	84.6%
3	16	12	75.0%

Table G.2: Model Performance on Web Forecasting

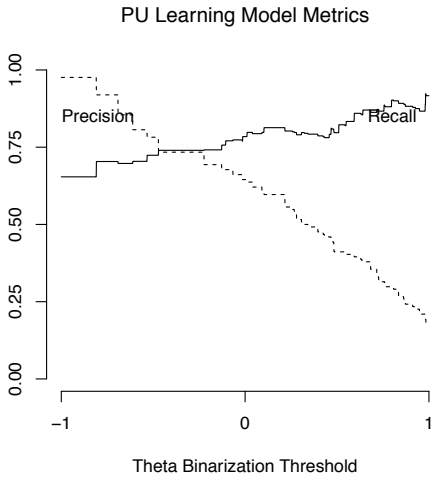
We now turn to our second evaluation test in which we compare classification results for 2017 with estimates derived from human experts. Figure G.1 presents the results for each of the three thresholds. In panels (a), (c) and (e), we compare the mean theta values (with 95% bootstrap CIs) for SIs that the three models predict as not significant and significant. The results from all the three models show a strong correspondence between human experts and our clasification. Figure G.1, panels (b), (d) and (f), report model precision and recall metrics for different binarization thresholds in the range between -1 and 1 on the theta scale. The one-rater model achieves best accuracy when one binarizes on theta values between -0.809 and -0.692 (Accuracy 0.73, Recall 0.70, Precision 0.92). The two-rater model achieves best accuracy when one binarizes on theta values between 0.103 and 0.216 range (Accuracy 0.70, Recall 0.67, Precision 0.64). The three-rater model has best accuracy when one binarizes on theta values between -0.809 and -0.692 (Accuracy 0.70, Recall 0.67, Precision 0.89). Overall, the accuracy for all three models varies between 0.70 and 0.73.

Finally, we replicate our third evaluation test in which we probe the construct validity of our estimates. Figure G.2 presents the results for each of the three thresholds. There are two main points to note. Figure G.2, panels (a), (c) and (e), all show a marked surge in the production of significant SIs in February and March, the two months before the budget. Second, Figure G.2, panels (b), (d) and (f), demonstrate that the significance of SIs, as captured by our classification, increases strongly as one moves closer to the budget date.

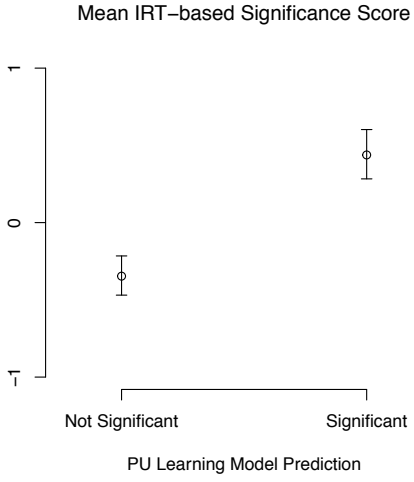
We conclude that our results show robustness to different threshold specifications.



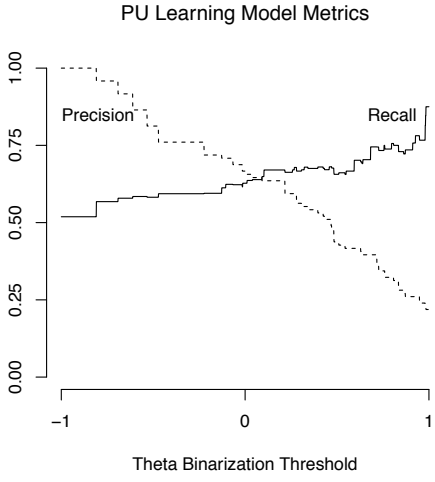
(a) One-Rater Threshold, $n = 1$



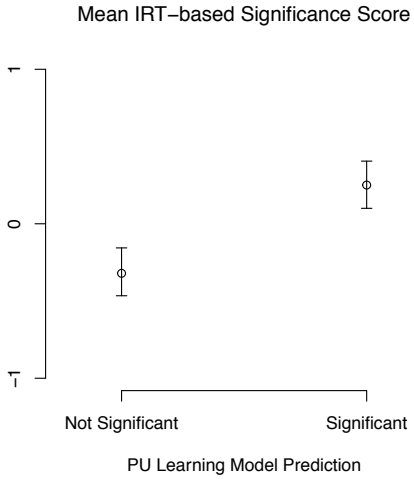
(b) One-Rater Threshold, $n = 1$



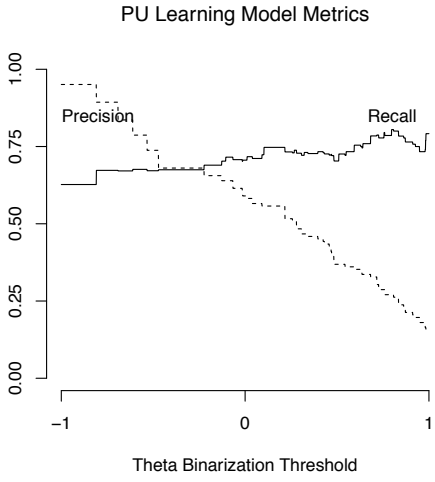
(c) Two-Rater Threshold, $n = 2$



(d) Two-Rater Threshold, $n = 2$

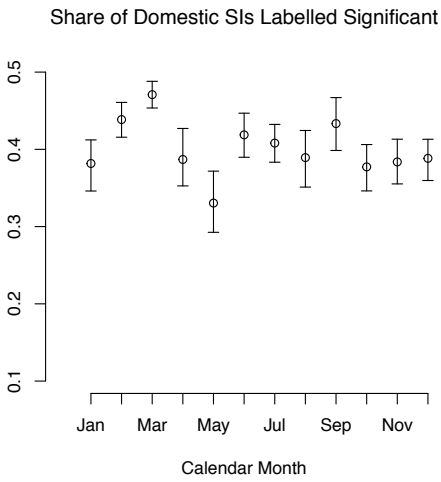


(e) Three-Rater Threshold, $n = 3$

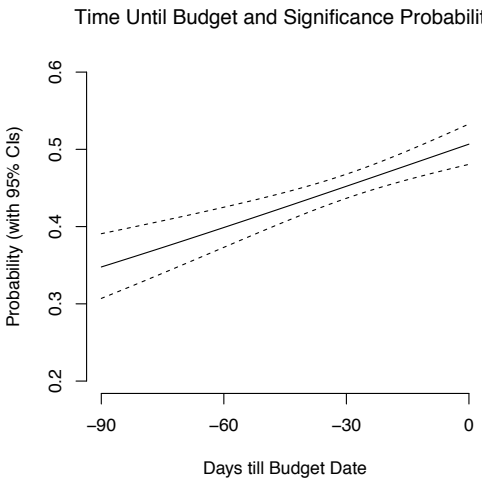


(f) Three-Rater Threshold, $n = 3$

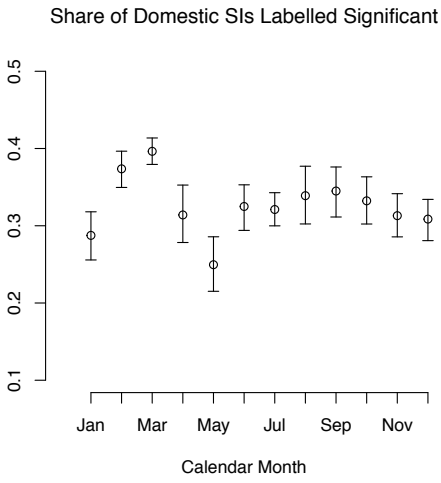
Figure G.1: Comparison with Human Raters



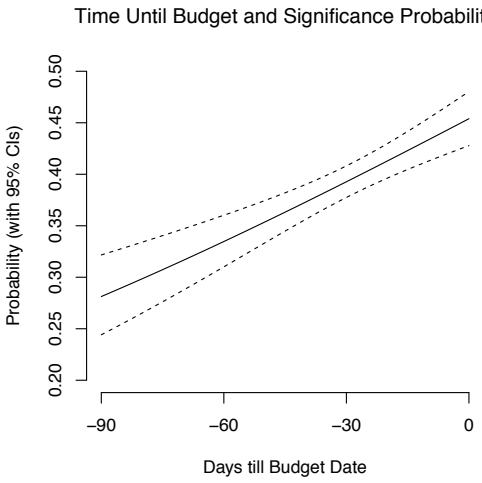
(a) One-Rater Threshold, $n = 1$



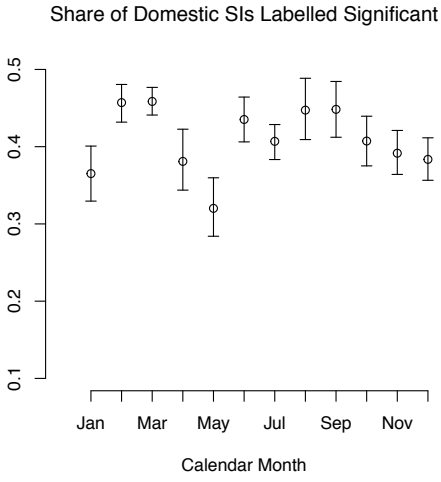
(b) One-Rater Threshold, $n = 1$



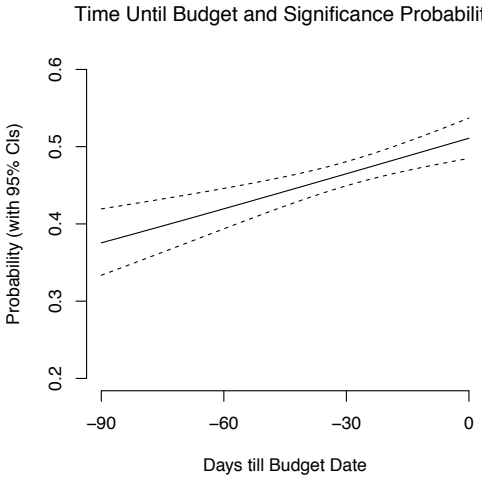
(c) Two-Rater Threshold, $n = 2$



(d) Two-Rater Threshold, $n = 2$



(e) Three-Rater Threshold, $n = 3$



(f) Three-Rater Threshold, $n = 3$

Figure G.2: Statutory Instruments and UK Budget Cycle

G.2 Analysis with Alternative Firm Sets

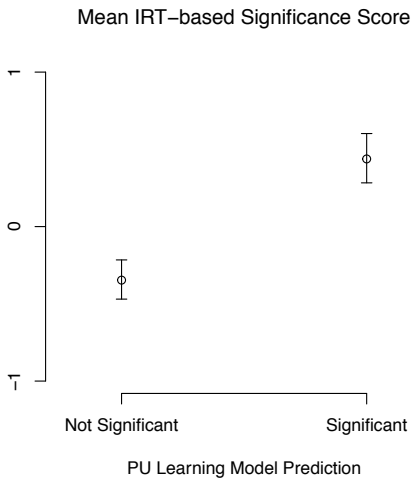
The three PU learning models discussed in the previous section are trained using the positive examples of laws which we obtained from Google searches performed on the websites of all 288 law firms. Would our results change in an important way if we were to vary the number of web pages we query? To address this point, we undertake two additional robustness tests in which we vary the number of firms from which we harvest examples of significant laws, while keeping the threshold constant at n = 2. The overall takeaway from this analysis is that our results remain substantively the same.

Table G.3 presents the key training-stage metrics for the models estimated with P_2 sets obtained from varying numbers of law firms. The P_2 obtained from 154 firms is our baseline scenario. While we query the websites of 288 law firms, our Google search finds one or more hits in the web pages of 154 firms. For the purposes of the robustness check, we replicate our analysis with two new P_2 sets obtained from smaller sets of firms drawn randomly – a set of 125 firms and a set of 75 firms. Table G.3 presents the results. All three models demonstrate a good model fit on training, with the true positive rate between 91.3 % and 93.6%.

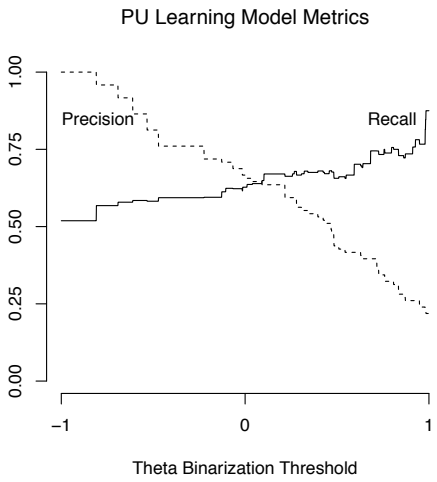
P Set	Firms	Actual Positives	True Positives	True Positive Rate
P_2	154	271	251	92.6%
P_2	125	196	179	91.3%
P_2	75	109	102	93.6%

Table G.3: Model Metrics on Training Data (Two-Rater Threshold, $n = 2$)

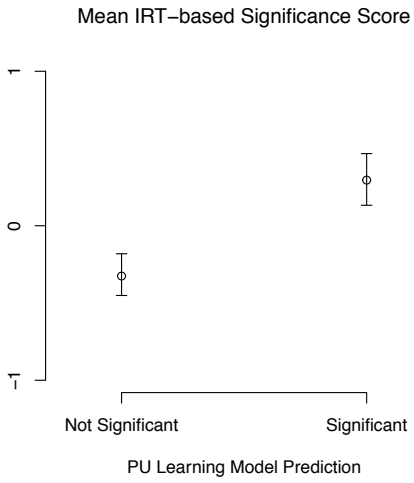
In evaluating performance beyond training, we first examine the metrics for forecasting citations within web content. Using the search results from all firms to create our test set, we find that out of the total of 26 SIs cited on the websites of at least two law firms, all of our three models correctly predict 22 positives, misclassifying four SIs. The true positive rate is thus 84.6% for all models regardless of the set of firms they have been trained on. Figures G.3 and G.4 present the metrics for the other two evaluation tests: a comparison with estimates derived from human experts and an examination of the construct validity of our estimates, respectively. The results are very similar across all the three models.



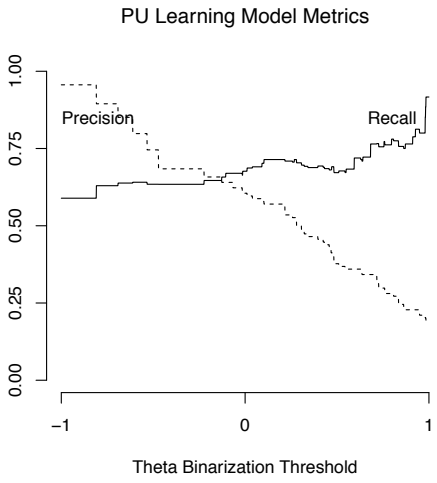
(a) 154 Firms, $n = 2$



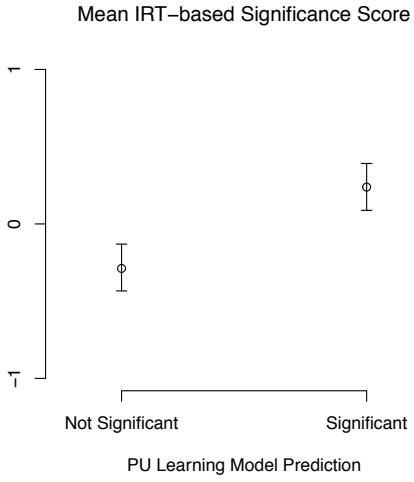
(b) 154 Firms, $n = 2$



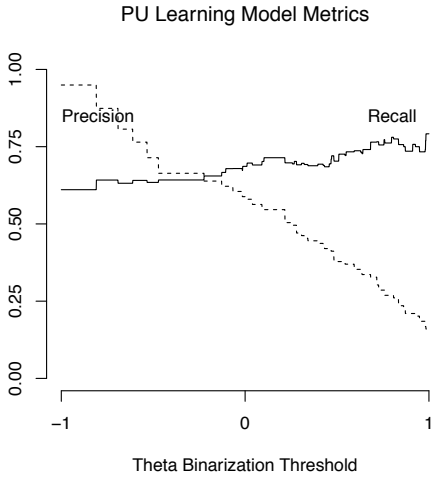
(c) 125 Firms, $n = 2$



(d) 125 Firms, $n = 2$

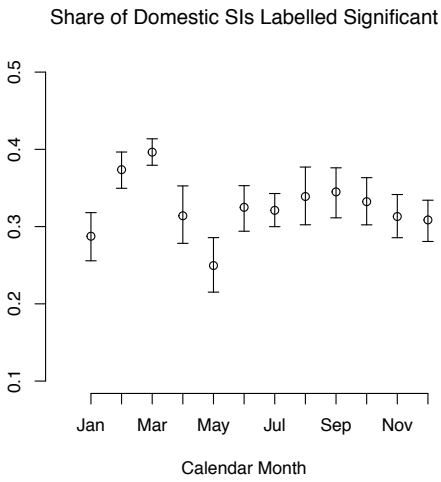


(e) 75 Firms, $n = 2$

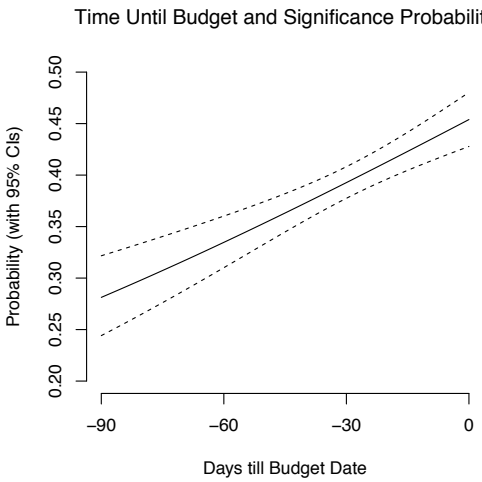


(f) 75 Firms, $n = 2$

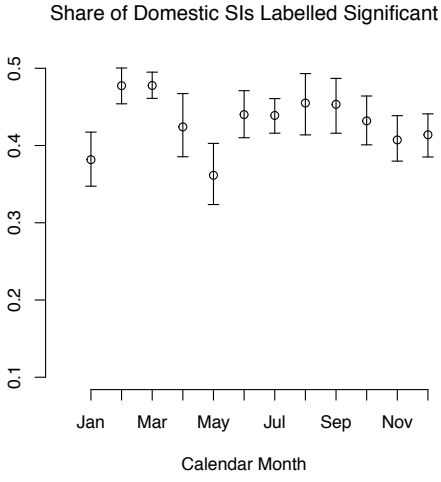
Figure G.3: Comparison with Human Experts



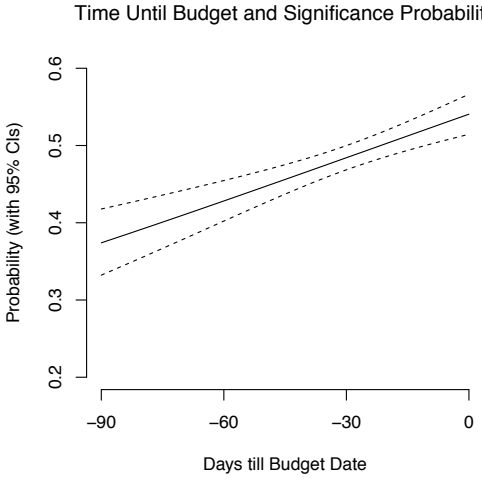
(a) 154 Firms, $n = 2$



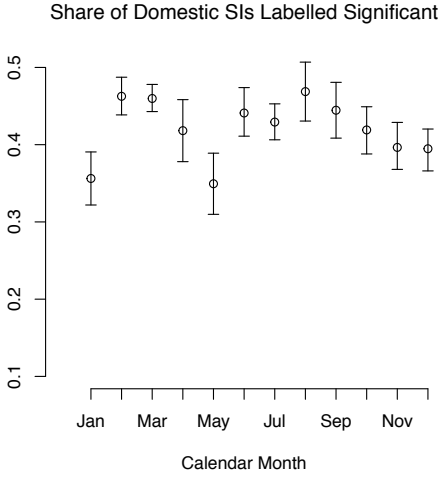
(b) 154 Firms, $n = 2$



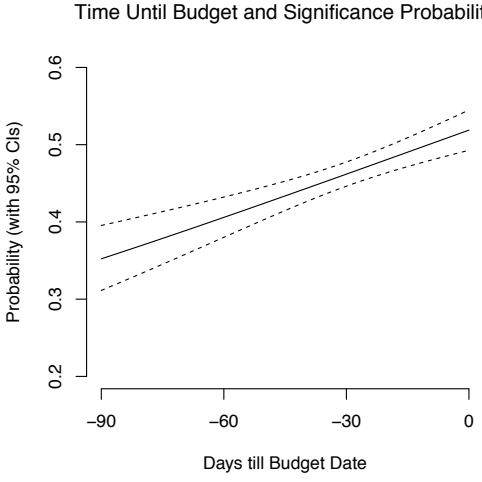
(c) 125 Firms, $n = 2$



(d) 125 Firms, $n = 2$



(e) 75 Firms, $n = 2$



(f) 75 Firms, $n = 2$

Figure G.4: Statutory Instruments and UK Budget Cycle

H Textual Feature Importances

While text features constitute a majority of our features (351,177 out of 353,592 features, 99.3%), individually, they have relatively less predictive power when compared to the categorical features. Nevertheless, some text features do have predictive value, and the top contributing and inhibiting text features are shown in Table H.1. The two top contributing features are 'introduction new' and 'new part', both of which signal the enactment of novel provisions. The top inhibiting features are more difficult to interpret, but the bigrams 'coming force', 'has not', and 'not produced' most likely signal the absence of additional documentation such as memoranda or impact assessments.

Table H.1: Importances of text features

contributing feature	mean rank
new part	183.78
introduction new	192.38
occupational pension	194.37
permitted development	204.77
provision relating	236.11
inhibiting feature	**mean rank**
coming force	167.90
has not	178.29
not produced	188.12
instrument no	193.92
upper tribunal	214.25

I Expert Survey

I.1 The Human Raters

To validate our estimates, we use ratings from three human experts. In May 2019, we recruited these experts by sending out a recruitment email to all law students at the University of Oxford. The three expert raters were chosen from a total of ten applications.

Our three expert raters were all graduate-level law students at the University of Oxford who all had some professional experience working at major international law firms when they completed our coding task. All three had previously completed undergraduate BA degrees in law: one at Oxford and two elsewhere (both in countries with legal systems akin to the UK). Two of our raters were current students on the Bachelor of Civil Law at Oxford; a one-year taught graduate course in law, designed for those with common law backgrounds. According to the University's website, for the BCL, "The academic standard is significantly higher than that required in a first law degree, and only those with outstanding first law degrees are admitted."[26] Our third rater had already successfully completed the BCL degree at Oxford and was a student on the Master of Philosophy (Law). As such, our three expert raters were highly successful graduate law students, all of whom had completed undergraduate law degrees and all of whom, at the time of our rating exercise, were pursuing graduate-level legal studies at Oxford.

I.2 The Task

We asked each expert to read and rate the significance of 217 different SIs. The sample is partially random as we included 26 SIs, which a new Google search performed in February 2018 reveals to have been mentioned by at least two law firms plus a random draw of 50% from the remaining 383 SIs. We administered the survey using the Qualtrics survey platform between July and September 2019. Each expert rater was sent a link to the survey. Once they clicked on this link, they were presented with the question: '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?' Below this question, was a hyperlink to the relevant SI and each rater then chose a rating from the following scale: 'Extremely Important', 'Very Important', 'Important', 'Moderately Important', 'Slightly Important', 'Not at all important'. A screenshot from Qualtrics illustrating the set-up of the task can be found in Figure I.1.

[26] https://www.ox.ac.uk/admissions/graduate/courses/bachelor-civil-law?wssl=1, accessed April 2020.

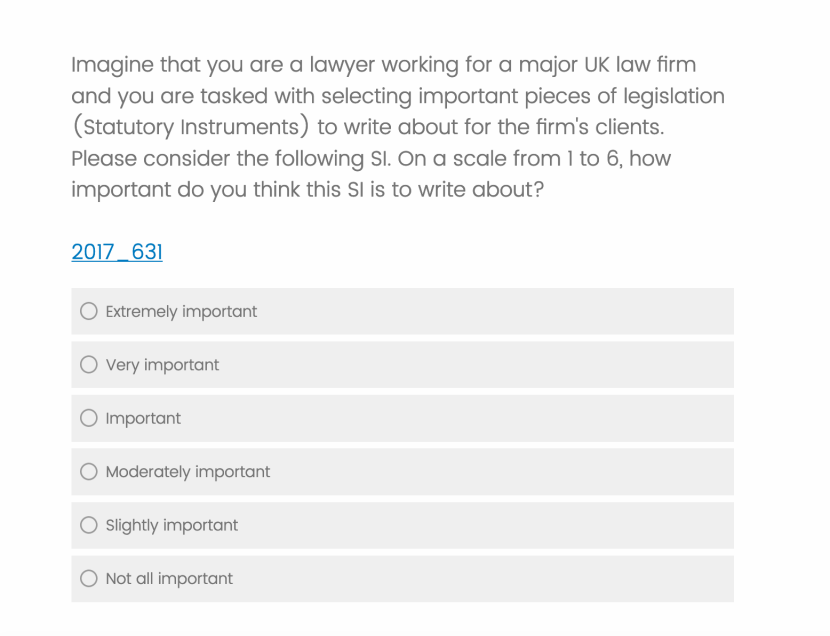


Figure I.1: Screenshot from Qualtrics of human expert rater survey.

I.3 Interrater Reliability

There are three main ways of assessing inter-rater reliability: 1) consensus estimates, 2) consistency estimates and 3) measurement estimates (see Stemler 2004). Consensus estimates are based on the assumption that our raters are able to come to exact agreement on how to apply the different importance scales to our sample of SIs. If our raters do come to exact agreement, then it could be said that they share a common interpretation of the construct and have a shared understanding of how to use the scoring rubric. Such a shared understanding normally arises from extensive training of raters as to how to interpret and use a particular scoring system (see Stemler 2004). We did not engage in extensive training of our raters. Rather, we asked them to assign scores to our sample of SIs based on their interpretation of the scale we developed. This means that using consensus estimates to assess interrater reliability is not the appropriate approach in this context.

In fact, when we look at how our raters employed the scoring rubric, we can see that they were not all using the numeric scale in the same manner. Table I.1 illustrates the distribution of scores on our 1 to 6 importance scale by each of our three raters. It is evident that Rater 1 is not using the scale in a similar fashion to Raters 2 and 3. The distribution of the scores

of Raters 2 and 3 is more similar.

Rating	Rater 1	Rater 2	Rater 3	Total
1	52	134	104	290
2	41	36	39	116
3	21	22	29	72
4	34	15	21	70
5	31	8	15	54
6	38	2	9	49
Total	217	217	217	651

Table I.1: Interpretation of the rating scale across our three human raters

Table I.1 reinforces the point that consensus estimates are not suitable in this context, given that we did not train our raters in the use of our importance scale. Our raters clearly did not come to a unified consensus as to how to apply this scoring rubric and as Stemler (2004, 3) points out: "consensus estimates can be overly conservative if two judges exhibit systematic differences in the way that they use the scoring rubric". This is reflected in one of the most common consensus estimates of interrater reliability, Krippendorff's alpha. The alpha coefficient displayed in Table I.2 falls within the fair to moderate range (on the Landis and Koch (1977) scale), which is what we might expect, given the lack of rater training.

	Coef.	SE	Lower CI	Upper CI
Krippendorff's Alpha	0.2732	0.0490	0.177	0.370

Table I.2: Krippendorff's alpha coefficient. Based on irrCAC package in R.

Consistency estimates on the other hand, allow for raters to have a different interpretation of the rating scale, provided that they are consistent in how they apply this scale within the parameters of their own definition. Consistency estimates provide a means to summarize the scores of raters, while allowing for systematic differences in how individual raters use and apply the scoring scale. Because of this, it is possible to have low consensus estimates of interrater reliability (as in Table I.2), but high consistency estimates of interrater reliability.

Table I.3 reports a simple correlation matrix for our three human raters. Two points

stand out quite clearly. Firstly, the scores of Raters 2 and 3 correlate reasonably highly. The scores of Rater 1, in contrast, are less strongly correlated with those of our other two raters.

	Rater 1	Rater 2	Rater 3
Rater 1	1.000		
Rater 2	0.3731	1.000	
Rater 3	0.3825	0.6307	1.000

Table I.3: Correlation matrix for three human raters

Probably the most popular consistency estimate of interrater reliability for multiple raters is the Cronbach's alpha coefficient. Cronbach's alpha is particularly useful to ascertain the degree to which a group of ratings underpin the measurement of a common dimension and if the coefficient is low, it indicates that the variance among the scores is due to error variance as opposed to actual variance among the scores. Table I.4 reports the Cronbach's alpha coefficient for our three ratings. As this table illustrates, our Cronbach's alpha coefficient is 0.7205, above the recommended threshold of 0.70, suggesting that our raters exhibit reasonably good consistency estimates of interrater reliability.

Item	Obs.	Sign	std.r	average.r	std.alpha
Rater 1	217	+	0.7307	0.6307	0.7736
Rater 2	217	+	0.8340	0.3825	0.5534
Rater 3	217	+	0.8379	0.3731	0.5434
Test scale				0.4621	**0.7205**

Table I.4: Cronbach's alpha, estimated with standardized items

I.4 Aggregation Methods

In order to aggregate the data from our human coders, we can turn to three different methods: a mean standardized rating; a principal components analysis of the ratings; or, item response theory (IRT) models. We chose a graded response IRT model to derive the theta values of latent significance for each SI from our human ratings (Clinton and Lapinski

2006). We also validate our data with the other two methods of aggregation in Section I.5 and the results remain the same. The IRT method is particularly useful to create a common measure among raters while accounting for rater heterogeneity, specifically when raters employ the same criteria, but apply different thresholds and discrimination rates when considering SIs to be significant or not (see Clinton and Lapinski 2006, 237). Among our raters, this is particularly characteristic of Rater 1 when compared to Rater 2 and 3.

Table I.5 reports the estimated discrimination and difficulty parameters for our three expert raters from the graded IRT model[27]. We can clearly see there is notable variation in the discrimination rates and the thresholds employed by our raters. Rater 1 has a considerably lower discrimination than either Rater 2 or 3. Rater 1 also has a lower threshold, as illustrated by the difficulty parameters. Rater 1 has a 50 per cent chance of choosing significance level 1 with θ at -1.378, as opposed to 2 or higher. This is notably lower than the difficulty parameters for Raters 2 and 3 for this category. Raters 2 and 3 clearly have a higher threshold for significance than Rater 1. This is exactly what we would anticipate, given the distribution of scores in Table I.1 and this is what our IRT model can account for.

We can also explore whether our raters' assessments reflect a common 'evaluative dimension' by estimating a two-dimensional IRT model to explore the dimensionality of our rater scores (see also Clinton and Lapinski 2006, 241). For this we use the graded response model option in the `mirt` R package with 2 dimensions. A model comparison shows a slightly increased AIC (1808.29, up from 1804.29) and BIC (1875.89, up from 1865.13) for the two dimensional model as compared to the one dimensional model. The slight increase in AIC and BIC show that it is preferable to select the one dimensional model. These results suggest that our raters understand significance in terms of a single latent dimension.

The results from the principal components analysis also indicate that our human experts understood the importance of SIs as comprising one single dimension. As Tables I.6 and I.7 illustrate, the data from the raters has one major underlying component, with an eigenvalue of 1.93, which captures 65 per cent of the variance. The eigenvectors in Table I.7 also

[27]We use the `mirt` R package formulation of the graded response model which does not multiply the discrimination a with the difficulty b, the probability being of the form $S(a \cdot \theta + b)$, rather than the standard IRT formulation of $S(a \cdot (\theta - b))$. Thus the difficulty parameters from the model have inverted signs compared to the standard formulation. Here, we have presented the threshold estimates from the model with an inverted sign to align with the standard. $S(x)$ is the sigmoid function $1/(1 + \exp(-x))$.

	Rater 1 b/se	Rater 2 b/se	Rater 3 b/se
Discrimination	0.993***	3.120***	2.884***
	(0.18)	(0.92)	(0.79)
diff >=2	-1.377***	1.099***	-0.169
	(0.19)	(0.40)	(0.29)
diff >=3	-0.344**	2.834***	1.392***
	(0.16)	(0.71)	(0.40)
diff >=4	0.135	4.255***	2.743***
	(0.16)	(0.99)	(0.62)
diff >=5	0.940***	6.061***	4.124***
	(0.18)	(1.40)	(0.88)
diff =6	1.857***	8.710***	5.937***
	(0.22)	(2.05)	(1.22)

Table I.5: Discrimination and difficulty parameters, (***) $p < 0.001$ (**) $p < 0.01$

demonstrates that all three human raters load positively on this dimension.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.9358	1.2407	0.6453	0.6453
Comp2	0.6951	0.3259	0.2317	0.8769
Comp3	0.3692	.	0.1231	1.0000

Table I.6: Principal Components Analysis

Variable	Comp1	Comp2	Comp3
coder1	0.4958	0.8683	0.0165
coder2	0.6126	-0.3632	0.7021
coder3	0.6156	-0.3380	-0.7119

Table I.7: Eigenvectors

Qualitative evidence further emphasizes that our raters shared a common theoretical understanding of legislative significance. When asked to consider what Statutory Instruments might be significant to write about for a legal blog or website, our expert raters appeared to adopt an understanding of importance that was rooted in the idea that important legislation shifts the legal status quo. In September 2019, we asked our expert raters to provide a short

qualitative rationale as to why they gave a small group of SIs the rating they did. Their responses and their rationale for their ratings were all based on the notion of deviations from the legal status quo. For example, when asked why they rated the 2017 *Export Control Order* as extremely important, one of our expert raters suggested it was because "this order amends the previous Export Control Order 2008 which has relevant consequences for client transactions covered by the previous order". Another of our expert raters, when asked why they rated the 2017 *Crime and Courts Act Order* as slightly important, stated: "This changes the rules applicable to investigations in England & Wales, as well as NI". Another expert rater, who considered this SI to be extremely important, noted their rationale for this choice: "...it brings forth key changes to the 2013 act especially in terms of procedures for investigation and discovery". Clearly, our expert raters, when we asked them to consider which SIs would be important to write about, based their ratings on the degree to which each SI would alter the existing status quo. It is reassuring that not only did they share this common understanding of importance, but also that this understanding reflected the more general interpretation of significance adopted in the paper.

I.5 Validation

To validate our results, and as discussed in the main body of the paper, we used the theta values from our graded response IRT model as the true (continuous) estimates of SI significance as viewed by our human raters. We then compared the mean theta values (with 95% bootstrap CIs) for SIs that our model predicts as not significant and significant in Figure 3(a). In Figure 3(b) in the paper, we also considered the precision and recall of our PU learning model by examining how precision and recall change as one varies the threshold from -1 to 1 on the theta scale. This analysis demonstrated that our model has best accuracy when one binarizes on theta-based significance values between 0.103 and 0.216 (Accuracy 0.70, Recall 0.67, Precision 0.64).

We replicate these validation analyses with two alternative methods of aggregation: a mean of standardized scores and principal component analysis. For our mean of standardized scores, we firstly standardize the scores of our three human experts with a mean of 0 and a standard deviation of 1. We then take the mean of these standardized scores as a continuous

estimate of SI significance as viewed by our raters. Figure I.2(a) and (b) then replicate the analyses in Figure 3, replacing theta with the mean of standardized rater scores. In Figure I.2(a), we compare the mean of standardized scores (with 95% bootstrap CIs) for SIs that our model predicts as not significant and significant. The mean of standardized significance scores for the SIs predicted as not significant is -0.34 and that for the SIs predicted as significant stands at 0.43. In Figure I.2(b), we consider how precision and recall change as one varies the threshold from -1 to 1 on the mean of standardized scores scale. This analysis demonstrates that our model has best accuracy when one binarizes on mean standardized score-based values between 0.045 and 0.085 (Accuracy 0.71, Recall 0.70, Precision 0.60).

For our third aggregation method, we perform principal components analysis and then generate scores for each rating based on the first component. In Figure I.2 (c), we compare these PCA-based scores (with 95% bootstrap CIs) for SIs that our model predicts as not significant and significant. The mean of PCA significance scores for the SIs predicted as not significant is -0.58 and that for the SIs predicted as significant stands at 0.73. Figure I.2(d) considers precision and recall. Our model achieves best accuracy when the PCA scores are binarized at the threshold of -0.014 to 0.087 (Accuracy 0.71, Recall 0.69, Precision 0.60).

In Table I.8 we also display the confusion matrix between our human raters and our classification. To generate this confusion matrix, we binarized our theta scores for our human ratings where we received the optimal level of accuracy based on Figure 3(b). All scores below or equal to the threshold of 0.103 – 0.216 were coded as zero, and all scores above this threshold were coded as 1. From Table I.8, our confusion matrix has: Accuracy = 0.70; True Positive Rate = 0.67; and True Negative Rate = 0.72.

		Our model		
		0	1	Total
IRT theta	0	91	35	126
	1	30	61	91
	Total	121	96	217

Table I.8: Confusion matrix with theta dichotomized at threshold between 0.103 to 0.216

Table I.9 replicates the confusion matrix in Table I.8 but here we use the mean of standardized scores from our human raters and binarized these scores at the threshold where

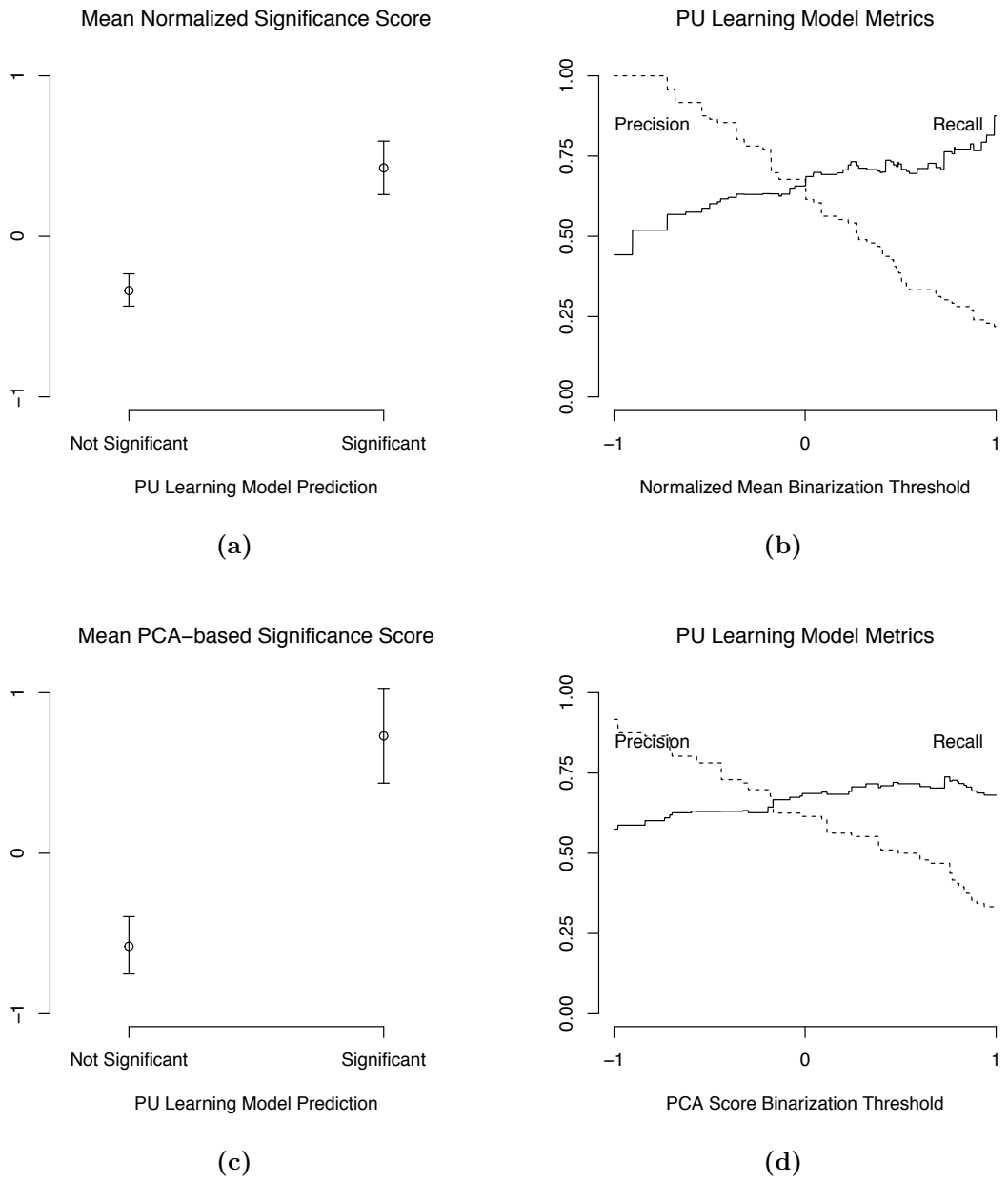


Figure I.2: Expert Validation with Alternative Rating Aggregation Methods

we obtain the best accuracy (as per Figure I.2(b)), where all scores below or equal to the threshold of 0.045 to 0.085 were coded as 0 and all scores above this threshold were coded as 1. In this instance, our confusion matrix has: Accuracy = 0.71; True Positive Rate = 0.70; and True Negative Rate = 0.72.

		Our model		
		0	1	Total
Mean z-scores	0	96	38	134
	1	25	58	83
	Total	121	96	217

Table I.9: Confusion matrix with mean standardized rater scores dichotomized at threshold between 0.045 to 0.085

Table I.10 reports the results of a third and final confusion matrix, but this time with the PCA-based scores of our human raters. Again, we binarize the data where we have greatest accuracy, with all PCA scores below or equal to the threshold of -0.014 to 0.087 coded as 0, and all scores above this threshold coded as 1. With this measure, our confusion matrix has: Accuracy = 0.71; True Positive Rate = 0.69; and True Negative Rate = 0.72.

		Our model		
		0	1	Total
PCA scores	0	94	37	131
	1	27	59	86
	Total	121	96	217

Table I.10: Confusion matrix PCA scores dichotomized at threshold between -0.014 to 0.087

I.6 Error Types

In this subsection, we undertake a brief analysis of classification errors made by our model relative to our human raters. We focus on the results presented in Table I.8. This is our main model and it is reasonable to examine the type of errors that this model makes. In Table I.11, we report the numbers of false negatives (FN), false positives (FP), true positives (TP) and true negatives (TN) broken down by the type of a statutory instrument.

	FN	TN	TP	FP	Total
All	30	91	61	35	217
Amendment	20	46	33	17	116
SI Type:					
Regulations	11	36	53	34	134
Orders	15	53	8	1	77
Rules	4	2	0	0	6
Long	12	20	40	20	92
Application:					
UK-wide	18	25	60	27	130
Regional	12	66	1	8	87
Commencement	6	6	0	0	12
Transitional	5	0	2	0	7

Table I.11: Error Analysis by SI Type

This table reveals that if we examine in more detail those SIs that our model misclassifies relative to our human raters, we can clearly observe a preponderance of regulations – of all 65 SIs that our model misclassifies relative to our human coders, 45 are regulations (69%). This same is true of SIs with UK-wide applicability which also make up 69% of all misclassified cases. Also, a large number of misclassified cases are amendments – 57 per cent. Finally, almost half of misclassified cases are SIs with above-average text length.

A closer inspection of individual errors suggests that our model tends to misclassify more nuanced cases. For example, amendment orders with below-average length would largely be classified as non-significant by both our model and human experts. But there are clearly some such SIs which human raters view as important. Consider *The Capital Allowances Act 2001 (Cars Emissions) (Amendment) Order 2017*, one of the SIs which our model predicts as negative but which human experts consider to be significant. On the surface, this SI appears appears simply to amend the existing *Capital Allowances Act 2001 (Cars Emissions) Order* but in fact, this amendment changes the emissions threshold for car lease restrictions for the purposes of income tax and corporation tax. This means that for corporate entities with company cars, this amendment would have consequences for their capital allowances. This

SI is a major deviation from the status quo.

Other examples of nuanced cases include new regulations with above-average length which, on the whole, both our model and human experts tend to classify as significant. But there are clearly some such SIs that experts do not consider as important, while our model classifies as significant. Consider *The Town and Country Planning (Brownfield Land Register) Regulations 2017*. This SI places a duty on local planning authorities to prepare and publish a register of previously developed land which is suitable for residential development. The subject matter is new and the regulatory requirements are specified in great detail. At the same time, the regulations introduce these new provisions only in England and only within the domain of internal local government procedures.

Our model struggles to pick up theses subtle, but very important nuances. Of course, our PU learning model is not unique in this respect. Automatic classifiers, while highly efficient at labelling, often struggle to discern important nuances that humans can easily interpret and observe (see Grimmer and Stewart 2013). We can only speculate about the source of our errors. There may be not enough training data for some important combinations of our features. Our model may also rely too heavily on categorical features, as our textual features come from texts that tend to use dry and bureaucratic language which reduces useful signal in the data. As our purpose here is to offer a proof-of-concept for a new approach, we leave the task of further improving model performance for future work.

J Forecasting Web Citations: Misclassified SIs

Table J.1 lists the four SIs, which appeared on the websites of at least two law firms in the first half of 2017, but which our model failed to predict as significant.

K Probability Calculation

Let the total number of SIs be N, out of which the number of SIs predicted as positives is P. The total true positive number is T. Of the predicted positives P a subset A are true positives. We want to compute the likelihood that A was chosen just by chance, given that

Table J.1: Misclassified positives

si	title	mean	a_i
2017_95	The Civil Procedure (Amendment) Rules 2017	0.00	0.00
2017_80	The Bank of England and Financial Services (Consequential Amendments) Regulations 2017	0.07	0.02
2017_402	The Town and Country Planning (Permission in Principle) Order 2017	0.11	0.05
2017_321	The Nursing and Midwifery (Amendment) Order 2017	0.41	0.29

we have chosen P out of N. Thus the denominator in the fraction is the number of ways to choose P out of N or ${}^{N}C_{P}$. Since A is a subset of P, the variation comes from the $P-A$ SIs which are predicted positive but not true positives. Keeping A fixed, we get the numerator then as the number of ways to choose $P-A$ out of $N-A$ times the number of ways to choose A out of T, the true positives: ${}^{N-A}C_{P-A} \times {}^{T}C_{A}$. Then the likelihood of having picked A by chance is

$$\frac{{}^{N-A}C_{P-A} \times {}^{T}C_{A}}{{}^{N}C_{P}} \tag{7}$$

We have $N = 409, P = 155, A = 22, T = 26$ and this gives us a very low probability of 2.98×10^{-6}.

L SIs and UK Budget

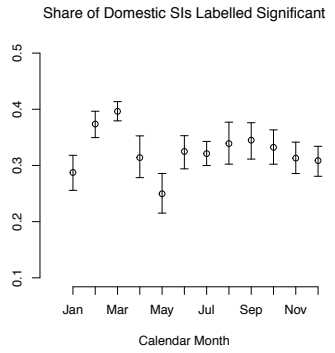
In this part of the Appendix, we check the robustness of our analysis of the UK legislative cycle by re-estimating it with alternative samples and model specifications.

In the main article, we limit our analysis to domestic SIs, that is, we exclude any SIs which were adopted in 2015-2015 to implement European Union laws.[28] As a member of the European Union (until recently), the UK was required every year to implement numerous EU directives and this process occurred mainly through the adoption of secondary legislation. The exclusion of EU-related SIs makes sense in principle, as such SIs are adopted subject to implementation deadlines externally determined in the EU law. As a first robustness check, we analyze how the significance of SIs varies over the annual legislative cycle if one uses the full sample of statutory instruments adopted in 2005–2010. We would expect our main effect to hold, but to become slightly attenuated, and this is exactly what we find. See Figure L.1 (c) - (d). As expected, we also find that EU-related SIs adopted in the weeks before the budget are no more significant than SIs adopted in other months - see Figure L.1 (e) - (f).

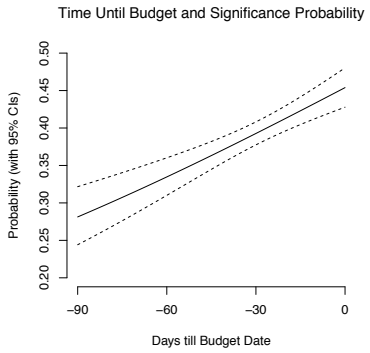
Our second robustness check is to truncate our SI sample to exclude election years. In 2005-2010, UK elections took place in May, and it is possible that the pattern we find in the UK legislative cycle reflects a pre-election boost in the production of significant legislation. To test whether this pattern still holds if one excludes election years, we remove SIs passed in 2005, 2010 and 2015 from our sample and re-run the analysis. Figure L.1 (g) - (h) reveals that our findings hold for non-elections years.

As a final robustness check, we re-estimate our GLM model from Figures L.1 (b), (d), (f) and (h) including year fixed effects. This analysis reveals that our expected effect of days until budget is robust to the inclusion of year fixed effects. The GLM results for all models are presented in Table L.1.

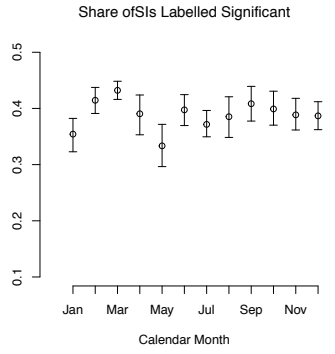
[28] We identified EU-related SIs by checking if the parent primary legislation for a given SI included the European Communities Act.



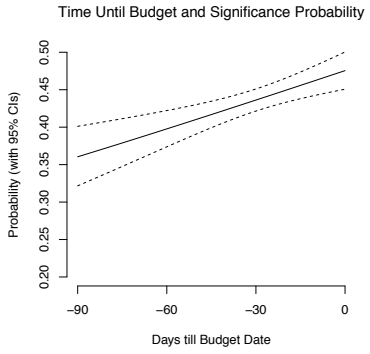
(a) Domestic SIs



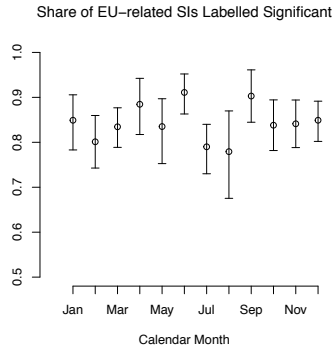
(b) Domestic SIs



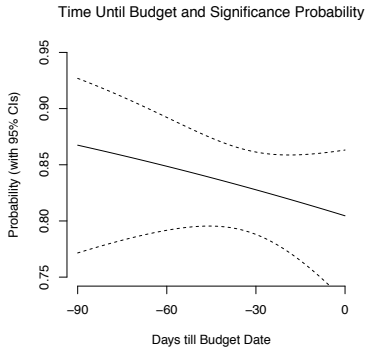
(c) All SIs



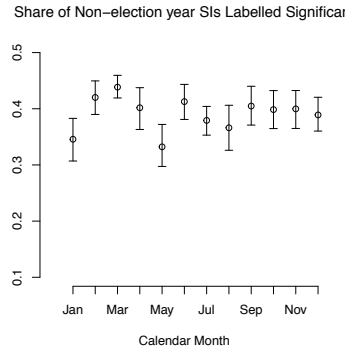
(d) All SIs



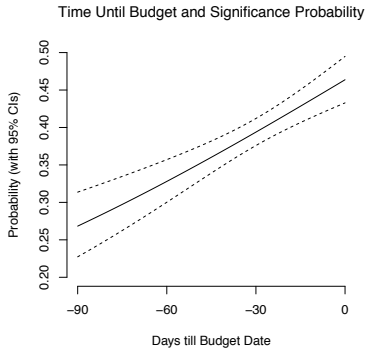
(e) EU-related SIs



(f) EU-related SIs



(g) Non-election Years



(h) Non-election Years

Figure L.1: Analysis with Alternative SI Samples

	Estimate	Std. Error	z value	p	
Domestic SIs					
Days till budget	0.008	0.002	5.566	0.000	***
(Intercept)	-0.185	0.054	-3.417	0.000	***
Domestic SIs (non-election years)					
Days till budget	0.010	0.002	5.488	0.000	***
(Intercept)	-0.145	0.064	2.280	0.023	*
Domestic SIs (with year FEs)					
Days till budget	0.009	0.002	5.789	0.000	***
(Intercept)	0.069	0.876	0.078	0.938	
All SIs					
Days till budget	0.005	0.001	3.858	0.000	***
(Intercept)	-0.098	0.051	-1.930	0.054	
EU-related SIs					
Days till budget	-0.005	0.005	-0.954	0.340	
(Intercept)	1.415	0.217	6.507	0.000	***

Table L.1: GLM Models Results. Significance indicators: (***) $p < 0.001$ (**) $p < 0.01$ (*) $p < 0.05$

References

Adam, Lucy. 2002. *Marketing your Law Firm: a solicitors' manual*. London: Law Society.

Barnett, Daniel, and Eugenie Verney. 2009. *Intelligent Marketing for Employment Lawyers: How to Boost Your Profits in a Recession*. London: ELS Publishing.

Brushfield, Rachel. 2018. *Smarter Legal Marketing: Practical Strategies for the Busy Lawyer*. London: Law Society.

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

Denis, François, Rémi Gilleron, and Fabien Letouzey. 2005. "Learning from positive and unlabeled examples". *Theoretical Computer Science* 348 (1): 70–83.

Fung, Gabriel Pui Cheong, et al. 2005. "Text classification without negative examples revisit". *IEEE transactions on Knowledge and Data Engineering* 18 (1): 6–20.

Furlong, Jordan, and Steve Matthews. 2014. *Content Marketing and Publishing Strategies for Law Firms*. London: Ark Group.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". *Political analysis* 21 (3): 267–297.

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

HMSO. 2006. *Statutory Instrument Practice: A manual for those concerned with the preparation of statutory instruments and the parliamentary procedures relating to them*. London: Her Majesty's Stationery Office.

John, Peter, et al. 2013. *Policy Agendas in British Politics*. Basingstoke: Palgrave Macmillian.

Landis, J Richard, and Gary G Koch. 1977. "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". *Biometrics* 33 (2): 363–374.

Lee, Wee Sun, and Bing Liu. 2003. "Learning with positive and unlabeled examples using weighted logistic regression". In *Proceedings of the Twentieth International Conference on Machine Learning, ICML*, 3:448–455.

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*. New York: Springer.

Liu, Bing, et al. 2003. "Building text classifiers using positive and unlabeled examples". In *Third IEEE International Conference on Data Mining*, 179–186. IEEE.

Manevitz, Larry M, and Malik Yousef. 2001. "One-class SVMs for document classification". *Journal of machine Learning research* 2:139–154.

Monk, David, and Alastair Moyes. 2008. *Marketing Legal Services: Succeeding in the New Legal Marketplace*. London: Law Society.

Page, Edward C. 2001. *Governing by numbers: Delegated legislation and everyday policy-making*. Oxford: Hart Publishing.

Rocchio, Joseph John. 1971. “Relevance feedback in information retrieval”. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, ed. by Gerard Salton, 313–323. New Jersey: Prentice-Hall.

Schölkopf, Bernhard, et al. 2001. “Estimating the support of a high-dimensional distribution”. *Neural computation* 13 (7): 1443–1471.

Smith, Nathan. 2014. *Social Media in the Legal Sector*. Londomn: Law Society.

Stemler, Steven E. 2004. “A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability”. *Practical Assessment, Research, and Evaluation* 9 (1): 4–15.

Zhang, Bangzuo, and Wanli Zuo. 2008. “Learning from Positive and Unlabeled Examples: A Survey”. In *2008 International Symposiums on Information Processing*, 650–654. IEEE.