Which laws are significant? Applying machine learning to classify legislation

Radoslaw Zubek, Abhishek Dasgupta, and David Doyle introduce a novel machine-learning approach to identifying important laws. They apply the new method to classify over 9,000 UK statutory instruments, and discuss the pros and cons of their approach.

Thousands of laws are published every year. In Britain, more than 300 public acts and almost 25,000 statutory instruments reached the statute book between 2010 and 2020. But which of these laws are really significant, and which ones are relatively minor? This is an important question for businesses and individuals. It is also one that many social scientists grapple with when studying law-making.

Conventional approach

The conventional approach is to ask experts – lawyers, reporters or policy professionals. The recipe is simple: find a group of reputable experts and ask them to classify a set of laws into those they find notable and those which they do not; in the final step, combine such individual evaluations into a total score using some aggregation method.

This is a great approach which has been employed with some success. But it is not without its problems. For one, it is time consuming and labour intensive. Perhaps more importantly, it also struggles to ensure that experts apply the same concept of significance and that they give equal weight to both recent and older enactments. How can we improve on it?

Our novel approach

In our recent article, we offer a proof of concept for a novel approach which we think has important advantages with respect to increased automation, reproducibility, and minimisation of recall bias. Our method has two major steps.

In the first step, we harvest 'seed' sets of significant laws from web data. A few billion people worldwide upload millions of posts every day on a myriad of issues including legislation. By posting content online, users signal which laws they consider significant. Also, many contributors, e.g., market analysts and law firms, are specialised domain experts. We take advantage of this propensity to freely share professional opinions.

In the second-step, we train a positive unlabeled (PU) learning algorithm. Recent advances in machine learning have offered sophisticated methods for building models when only positive examples are available, including two-step methods, biased two-class classifiers, and one-class classifiers. We employ PU learning to construct a computational formula that finds laws that are similar to our 'seeds' (positives) within a large pool of unlabeled legislation.

Application: UK Statutory Instruments

We apply this approach to classify UK statutory instruments, the most common (and most plentiful) form of secondary legislation in the UK. In our application, we source examples of significant laws from the web pages of top-ranked UK law firms. Websites offer an attractive platform for law firms to demonstrate expertise within their practice areas. Regulatory updates drawing attention to important changes in legislation are a key part of these marketing activities. We perform an automated search of the websites of 288 leading law firms and obtain a set of 271 important instruments.

We train our model using an adapted version of an established two-step Rocchio-SVM method. Our training data consists of web-sourced positives and a set of all UK statutory instruments adopted between 2009 and 2016. To train the algorithm, we rely on two types of information: textual features obtained from explanatory notes and a battery of categorical features such as topic, department, and length.

A key test for our model is whether it is able successfully to predict outside the training data. We evaluate our approach in three ways. We first check if our model is able to predict future law citations on the web, and we find a high true positive rate of 85%. We then compare our automated classification with hand-coded ratings, and we achieve a fairly high accuracy of 70%. Finally, we examine how the share of laws we classify as significant varies over the annual legislative cycle in the UK and we find that our method produces estimates with high construct validity. All in all, we think our method shows good promise.

Pros and cons

Our approach has clear advantages. Automation saves time and labour, and enhances reproducibility of classifications. We can also be specific about our definition of significance. In our application, we show that lawyers post content online mainly about laws that change the regulatory status quo by a large margin. With our web-based approach, we are also able to minimise recall bias by focusing on contemporaneous evaluations that assess significance of laws around the time of their enactment.

Our method is not without its limitations, of course. 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. We leave the task of further improving our model performance for future work.

Note: the above draws on the authors' published work in *American Political Science Review*.

About the Authors

Radoslaw Zubek is Associate Professor in the Department of Politics and International Relations at the University of Oxford.

Abhishek Dasgupta is a Research Software Engineer in the Department of Computer Science at the University of Oxford.

David Doyle is Associate Professor in the Department of Politics and International Relations at the University of Oxford.

Photo by Fotis Fotopoulos on Unsplash.