Automated Measurement of Policy Similarity in Legislation: Revealing Patterns of Text Reuse

abstract start

Much political science research relies on identifying the substantive similarity of legislation, especially in the study of policy diffusion. Scholars mostly rely on hand-coded datasets. However, in order to study policy diffusion more broadly, for example by extending the study to dozens or more policy domains, prohibitive volumes of legislative text would have to be analyzed manually. Common text analysis algorithms such as topic models or bag-of-words based text similarity measures are too coarse to identify policy similarity. However, legislators often directly use text from bills introduced in other legislatures or bill text provided by interest groups in model legislation. We propose the use of text-sequencing algorithms to find matching text between bills. We describe a new approach to the application of sequence alignment algorithms to large amounts of legislative text, and assess the validity of text reuse as a measure of substantive bill similarity. Three key results, drawn from an analysis of 500,000 bills from US-state legislatures, demonstrate the validity of text reuse as a measure of policy similarity. First, we show that bills introduced by ideologically similar sponsors are more likely to exhibit a high degree of text reuse. Second, we show that the main topical themes underlying strings of reused text map closely onto major areas of public policy in the US States. Third, we show that rates of text reuse across state borders correlates strongly with the state policy diffusion ties data recently introduced by Desmarais, Harden and Boehmke (APSR, 2015).

abstract end

Sequencing algorithms, namely the Smith-Waterman local alignment algorithm, originate from DNA sequence matching and has been applied to text in the past, mostly for plagiarism detection (**??**). In political science, it has been used to track the legislative history of congressional bills (**?**). There are two major challenges for the application of the local

alignment algorithm to detect policy similarity: First, the algorithm does an exhaustive search for all potential matches between two text sequences, and it is therefore computationally unfeasible to calculate all pairwise similarities for large collections of bills. Second, a lot of text that is reused in bills is not of substantive interest. This boilerplate text, bill headers and footers and other procedural text, as well as reoccurring definitions or policy area specific phrases, is not indicative of bills having similar content. **?** tackled these problems by restrictively preselecting text that is analyzed with the local alignment algorithm and through selection of non-boilerplate text using supervised machine learning.

We propose a novel way to approach these problems. Our approach does not require the manual coding of text for the supervised classification of boilerplate text, and is less restrictive in the pre-selection of bills. Specifically, we apply a 3-step procedure that reduces the computational load as well as the amount of boilerplate text that is detected by the algorithm. First, we preselect bills that are analyzed with the alignment algorithm with a measure related to cosine similarity. After this we remove all text with a tf-idf score lower than a certain threshold to avoid analyzing clear boilerplate text and then apply the algorithm to this reduced data.

The second part of our analysis then focuses on how well the alignment between bills captures substantive policy similarity. In a first step we apply topic models to the identified text alignments in order to check whether the reused text captures substantively relevant policy language, or boilerplate text. If a clear boilerplate topic emerges we can quantify how large a portion of the identified alignments is policy relevant compared to boilerplate.

Furthermore, we use a hand-coded dataset of equivalent bills to investigate how well the local alignment algorithm predicts actual policy equivalence. In order to evaluate the validity of the text reuse measure on a basis broader then would be possible with hand-coded ground truth datasets, we assess the predictive validity of the measure. We do so in two ways: First, we asses how the measure performs in predicting policy diffusion

ties between states as established by **?**. If two states have a strong diffusion tie, higher numbers of aligned bills would be expected. And second, we assess how well the similarity measure on the bill level predicts the ideological distance between the bills' sponsors using state legislators' ideal points (**?**). That is, if two bills are sponsored by ideologically close legislators, we expect more text alignment than between bills of ideologically close legislators.

Our analyses show that the local alignment algorithm in combination with our pre and post processing procedure produces valid measures of policy similarity between bills. We aim to publish a dyadic dataset of similarity measures in order to allow a broad range of political scientists to use them for substantive research on US state politics.