Text classification of ideological direction in judicial opinions

Author(s):
Hausladen, Carina I.; Schubert, Marcel H.; Ash, Elliott

Publication date:
2020-06

Permanent link:
https://doi.org/10.3929/ethz-b-000406465

Originally published in:
International Review of Law and Economics 62, https://doi.org/10.1016/j.irle.2020.105903

International Review of Law and Economics 62 (2020) 105903

Contents lists available at ScienceDirect

International Review of Law and Economics

Text classification of ideological direction in judicial opinions

Carina I. Hausladen [a,b,*], Marcel H. Schubert [a,b], Elliott Ash [c]

[a] Max-Planck Institute for Research on Collective Goods, Germany
[b] University of Cologne, Germany
[c] ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Received 31 December 2019
Received in revised form 19 February 2020
Accepted 22 February 2020
Available online 28 February 2020

JEL classification:
C8
K0

Keywords:
Judge ideology
Circuit courts
Text data
NLP

ABSTRACT

This paper draws on machine learning methods for text classification to predict the ideological direction of decisions from the associated text. Using a 5% hand-coded sample of cases from U.S. Circuit Courts, we explore and evaluate a variety of machine classifiers to predict "conservative decision" or "liberal decision" in held-out data. Our best classifier is highly predictive ($F1 = .65$) and allows us to extrapolate ideological direction to the full sample. We then use these predictions to replicate and extend Landes and Posner's (2009) analysis of how the party of the nominating president influences circuit judge's votes.

1. Introduction

In the United States, judges wield significant power due to the common law system (Dainow, 1966). The extent of U.S. judges' influence is a motivation for the extensive research into the determinants of judicial decision-making. In particular, there is a large literature on how opinions are affected by the ideology of the respective judge (Segal and Cover, 1989; Martin and Quinn, 2002; Martin et al., 2004, e.g.).

A leading paper in this literature is Landes and Posner (2009). This paper looks at how the party affiliation of U.S. Circuit Court judges affects the political ideology of their votes (conservative or liberal) on the court. While judges are nominally non-partisan, party affiliation can be proxied by the party of the appointing president or the party share in the Senate at the time of appointment. Landes and Posner show that judge party affiliation is statistically related to the ideological direction of votes.

For their empirical analysis, Landes and Posner (2009) draw upon the Songer database of U.S. Circuit Courts,[1] which provides rich metadata, e.g., the political ideology of votes for each judge in each case. The classification of votes by ideological direction was a labor-intensive exercise which has led to frequent use in the empirical legal studies and political science literatures (Ginn et al., 2015; Reid and Randazzo, 2016; Landes and Posner, 2009, e.g.).

Notwithstanding its broad use in the literature, the Songer database has some limitations. First, the political ideology classification has been assigned by human coders, which could be error-prone. These errors add noise to regressions and complicate replicability. In particular, as noted by Landes and Posner (2009), the political positions of conservative/liberal are not constant over time. Therefore, data coded in the past may not be categorized correctly, and Songer Project ideology labels for older Circuit Court opinions may be systematically incorrect.

Another problem with the database is the sampling approach. First, the database is only available for 1925–2002, so empirical analysis of vote ideology is only possible for that time period. Second, only a small set of cases was labeled (just 5 percent of the cases for those years). Finally, the authors used stratified sampling to get labels for similar numbers of opinions across courts and time. Therefore, the dataset is not representative of the full distribution of circuit court cases.

The goal of this paper is to address these shortcomings using machine learning and natural language processing techniques. The idea is to treat a machine to code the ideological direction of the votes. Within the set of labeled case, we can check how well the algorithm replicates human labels.

* Corresponding author.
E-mail address: hausladen@coll.mpg.de (C.I. Hausladen).

[1] The original, as well as the extended versions, are available at songerproject.org.

https://doi.org/10.1016/j.irle.2020.105903

The classifier would provide a number of benefits. As soon as the classifier is trained, predictions even for an extremely large sample cost very little relative to hand-labeling (which require a human to read an opinion). We could potentially take the classifier to cases before 1925 and after 2002. Within the 1925–2002 period, we could classify the other 95 percent of unlabeled cases. Besides producing new labels, it could be used to audit and check existing labels for probable errors.

In this paper, we produce such a model. For the sake of interpretability, we focus on linear models. The model which worked best in our setting is a Ridge Classifier. Our model is trained on the complete opinion text in combination with the circuit, year as well as case type data. After optimization it achieves a cross-validated accuracy of 61.5% on the three label input and 66.5% on the two label subset. The final calibrated classifier working on the tow-label subset achieves the same accuracy score while increasing its precision as well its recall on the test set to 71.1% and 72.4% respectively.

With a validated data set in hand, we use it to undertake an extended replication of Landes and Posner (2009). First, we do our best to replicate the original paper and, despite some problems in replicating the original dataset, we could replicate significance as well as the direction of the most important coefficients. We extend the results and probe their robustness to multi-way clustering, group, and additional covariates. Finally, we show that the results hold partly when using our machine-predicted ideological labels as the outcome.

This paper contributes to the emerging literature applying data science techniques to empirical legal research questions. We review some of that literature in Section 2. After that, in Section 3 we describe the supervised learning task to predict ideological labels in circuit court decisions. Next, Section 4 reports the results of our replication study. Section 5 concludes.

2. Literature

This research sits at the intersection of two literatures. On one side, our paper is related to the research on judge ideology, which is focused on the positioning judges, mostly for the U.S. Supreme Court (e.g. Giles et al., 2001; Epstein and Segal, 2005; Epstein et al., 2012; Johnson et al., 2011; Kassow et al., 2012; Martin and Quinn, 2001; Masood and Songer, 2013; Ginn et al., 2015; Sturm and Pritchett, 2006; Randazzo et al., 2010; Reid and Randazzo, 2016).

The judge ideology literature has taken two main approaches. The first approach is to hand-coded cases by ideological direction. These include the Spaeth database for the Supreme Court and the Songer database for the Circuit Courts (Epstein et al., 2012; Sturm and Pritchett, 2006; Martin and Quinn, 2001; Epstein and Segal, 2005; Giles et al., 2001, e.g.). The second approach is to use a latent factor model based on the voting behaviour, to estimate a latent dimension for ideology based on judge agreement. This approach can identify median judges and the relative judge positioning on a scale over time (Martin and Quinn, 2002).

The advantage of the first approach is that the scale is interpretable, exists on the case level, and relies on expert judgment. However, it is costly and there are errors in coding. The advantage of the second approach is that it is cheap to compute for all judges, but it is not directly interpretable and does not exist at the case level. It also requires that judges vote in panels.

Our approach is something of a compromise, as we can form predictions for all cases and judges cheaply. It requires at least some hand-coding, but then can be applied to all cases. Methodologically, it is different because it uses the directly interpretable ideological labels of the hand-coded database. It does not assume a latent factor model, like Martin-Quinn. It also does not rely on contrasting votes of judges in a panel. This is relevant in our context because the large majority of decisions on the Appellate Courts do not have dissents. Voting behaviour is not necessary, only some hand labels and the original opinion text.

The second literature to which we contribute is that on using texts as data for social science research. In particular, to produce measures of ideology or partisanship. In law, an old study in this vein is Segal and Cover (1989), who use texts from newspaper editorials as a proxy for the ideology of newly appointed Supreme Court judges. More recently, popular methods in political science for scoring ideology in text include Wordscores (Laver et al., 2003), Wordfish (Slapin and Proksch, 2008), and Wordshoal (Landerdale and Herzog, 2016). These tools use statistical differences in word frequencies by topic. They are most useful for text corpora for which differences in ideology come through in different words. As opinions of (lower) judicial courts are constrained in their (permitted) wording opinion texts may only satisfy that criterion in a very limited fashion.

In the legal domain, our paper is most closely related to literature predicting case type (Undavia et al., 2018; Sulea et al., 2017; Boella et al., 2011) as well as that concerned with dimensions in judicial texts (see for example Ash and Chen, 2018; Ash et al., 2018). The three papers closest to ours, in goal as well as methodological approach, are by Landerdale and Clark (2014), Aletras et al. (2016), and Cao et al. (2018). In Landerdale and Clark (2014), the authors use an LDA model to estimate how different issues at stake in cases are related to Supreme Court judges' voting behaviour. The paper by Aletras et al. (2016) looks at decision direction of the European Court of Human Rights (ECHR) in regard to the violation of specific articles. The third paper, Cao et al. (2018) separates opinion texts into ideological and fact-driven parts and look at how well these different paragraphs predict case directionality. However, none of the approaches in those three papers are viable for our goal or data set. Landerdale and Clark (2014) use the underlying text but their focus on votes means that the approach is not applicable. In the case of Aletras et al. (2016), in a modelling perspective the approach is similar. However, their results rely on very clean data resulting in very homogeneous directionality criterion. As a consequence, it is more than a simple question of transferring their results. Last, the paper by Cao et al. (2018) does look at ideological directionality. The focus on paragraphs, however, means that an additional labelling effort is needed while we seek to minimize the costs of classification.

To recap, our paper contributes in the technical literature to the understand how to best implement a machine learning approach in the domain of judicial opinions. We aim to decrease labelling cost and increase scalability and reproduciblity compared to the hand-labelling approach while at the same time improving explainability relative to the latent modeling approach.

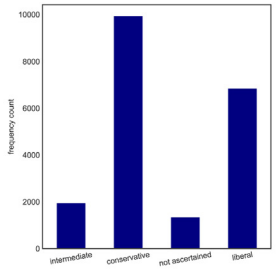
3. Supervised classification

This section focuses on the classification algorithm which can reliably predict the political ideology of Circuit Court judges' written opinions. After training the algorithm on existing ideology labels, it can predict labels for unseen opinions.

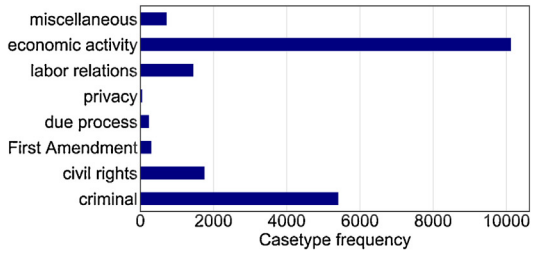
The beginning of this section provides information about the data necessary for classification. What follows is a detailed description of how the classifier is trained. Finally, the classification performance is evaluated.

3.1. Data

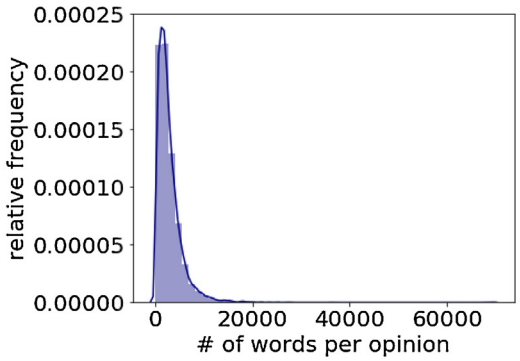
Broadly speaking, a supervised machine learning classifier maps an input to output. This section enumerates the datasets used for the inputs and outputs in our context. For our classification problem, we use the hand-coded ideology labels for these cases, provided by the Songer Project, as output. As input we use the U.S. Circuit Court judges' written opinions.



(a) Distribution of Political Ideology Labels



(b) Distribution of Case Types



(c) Distribution of Opinion Length (in Words)

Fig. 1. Summary statistics.

3.1.1. Songer data on decision direction

The output or label of our classifier is the ideological direction of the opinion. As the number of Circuit Court judges' opinions is over 300 thousand, the Songer Project has annotated political ideology labels for only a small sample of opinions, equalling less than 2.6% of the total published opinions available. The total is 769,986 when only taking those not decided per curiam into account. The cases were decided between 1925 and 2002 and the database contains a total of 20,355 cases. Overall, four directionality codes are available: "liberal", "conservative", "mixed" and "not ascertained". While "mixed" refers to the opinion of the case being of unclear directionality, "not ascertained" signals that the coders were unable to assign a label according to the codebook's instructions. Please note that directionality is defined for each particular case type, with "conservative" and "liberal" being exactly opposite outcomes. Fig. 1a shows the distribution of labels for the complete data-set. The categories "conservative" and "liberal" dominate, whereas the other two categories are underrepresented.

The Songer coders assigned the directionality of a case according to specific rules within case type. The case type of an opinion identifies the nature of the conflict between the litigants. Over 220 case type categories are organized into eight major categories: criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity and regulation, as well as miscellaneous. Fig. 1b shows the distribution of the eight major categories for our data-set. "Civil rights" and "economic activity and regulation" are the two case types most frequent in the data.

Landes and Posner (2009) mention in their paper that they applied substantial corrections to the raw Songer data, but those are not laid out in sufficient detail to reproduce. We approached the authors with the request to provide us with their version of the data-set. Unfortunately, they were not able to provide it yet.

3.1.2. Judicial opinion corpus

We matched the Songer data-set with the Lexis data-set, containing the full opinion text. With this approach, we could match 20,052 opinion texts to the 20,355 entries that the Songer database comprises. Regarding the non-matchable cases there is no clear pattern visible as these cases span nearly the complete time period as well as nearly all circuits. The distribution across time and circuits does not reveal any peculiarities either.

In terms of the matching itself, we subsetted the data according to the different circuits. That was only done for speed, as matching is a linear searching process which has to be repeated for each query. The actual matching was then done on either federal reporter citation or docket number. First, we tried to match via the normalized Lexis id, i.e. the Federal Reporter citation, if the opinion spanned more than one page in the Federal Reporter (to avoid confusion with other opinions). If such a match was not possible, we matched via the circuit court and the docket number. The reason why we preferred the federal reporter citation over the docket number is that the Songer database uses only encoded docket numbers. While they should be systematically encoded errors often result in decoding being little more than guess work. In the case of the federal reporter citation, errors were less prone.

Fig. 1c shows the distribution of opinions' word counts in our dataset. The shortest opinion consists of one word, the longest of 69,320 words. The average opinion consists of 2809 words. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single newline character. Special characters such as "newline"-characters and roman numbers were removed.

If a potential heading was found within the text, we excluded it. The reason being that such a heading would potentially include biasing information such as judge names. It is especially important to exclude those, as the model could focus on judge names as a proxy for the directionality as most cases were decided without dissent. This is an issue in our empirical context because we would like

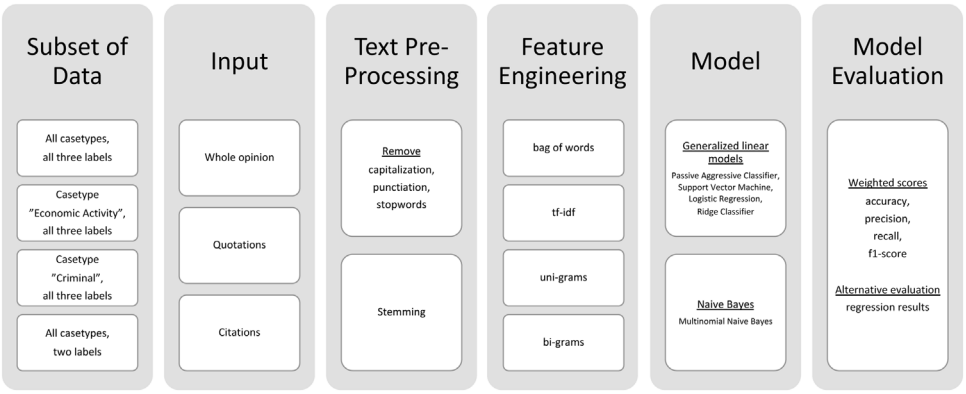


Fig. 2. Construction of the methodological approach.

to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions trying to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part which by its nature goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to an overall worse performance. If we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as dissent. We excluded those entries and did not use them as input.

3.2. Model

This section describes how we deploy a supervised learning approach to predict the ideological direction of decisions from the association opinion text.

Our approach, outlined by Fig. 2, is quite uncommon in the literature of classifying a legal text's ideology. More traditional approaches, mainly used for ideology detection in political speeches, include word scores, word fish, or word shoal models. These approaches are either dictionary-based or require a reference text to which all other instances are compared. Our approach, by contrast, does not require one reference text to be selected and deploys more sophisticated selection mechanisms than naive word counts.

One characteristic of machine learning approaches is their exploratory nature. We, too, test multiple combinations of data-subsets, feature sets, models, and evaluation methods to find the best performing one. The instances to test are either selected by theoretical considerations, such as choosing only judicial quotations as predictive features; Or they are chosen based on popularity, such as choosing support vector machines because they are known for their excellent performance on a broad range of NLP classification tasks.

All calculations were performed on the Max Planck Computing and Data Facility's high-performance cluster Draco, using one node of the type Broadwell with up to 40 CPUs and 256GB memory. Moreover, each step relying on randomness was initialized with a pseudo-random seed for replicability. Our code most heavily draws upon functionalities provided by the python package sci-kit learn (Pedregosa et al., 2011).

3.2.1. Subset of data

In order to see how different categories or a differing number of labels affects a prediction, we constructed different subsets of the data for analysis. Four subsets constructed from the original data and used for this analysis are listed in the first column of Fig. 2. A naive approach predicts political ideology labels regardless of case type. However, the naive approach ignores the fact that directionality in the Songer data is assigned dependant on case type according to explicit rules differing for each case type. Subsetting the data by case type factors in this aspect of the coding scheme.

However, as Fig. 1b shows, the data set is heavily imbalanced in favour of the case types "economic activity" and "criminal". As the remaining case types are only marginally represented, we restrict the subset two these two case types, as only for them enough labelled observations to train the classifier are available.

Moreover, not only case type but also the labels are imbalanced. As Fig. 1a shows, there is only a limited amount of observations available for the political ideology labels "not ascertained" and "mixed". We therefore derive two additional subsets. The subset "two labels" only includes the labels "conservative" and "liberal" as those two are not only the most frequent ones but also those we are most interested in. Especially if the remaining two labels ("not ascertained" and "mixed") are either considered as noise or wrongly classified, this subset should improve the classifier's performance. In particular, the exclusion of the label "not ascertained" is likely to not be problematic in any case: The number of cases labelled such are relatively few when compared to the other three labels. Moreover, the codebook shows that this label may be used in any case where it was not possible to assign one of the other three labels. This may either be due to the fact that the case truly fits into no other category or merely due to a lack in inter-coder agreement. However, past results show that such a sparsely represented, miscellaneous category decreases classification performance. For this reason, the final subset excludes this category altogether.

3.2.2. Input

We experimented with four different representations of the input. The most straightforward approach is to feed the complete pre-processed opinion text into the model. After screening a sample of randomly drawn opinions and cross-referencing them according to the labelling instructions from the codebook, we identified two additional representations.

First, we separately extracted the citations from the cases. The topic, as well as the political directionality of a case, might be captured already by citations. Citation networks, for example used by the *Supreme Court Mapping Project*, is one example using this reasoning (Chandler, 2005; Ash et al., 2018).[2]

Second, we extracted quotations from the text to serve as input. Many quotations immediately preceded citations. It is in the nature of a quotation that it represents the most relevant aspects to a

[2] see SCOTUS Mapper Library by the University of Baltimore.

matter at hand. As judges quote legal concepts from statutes and precedents relevant to the matter discussed, quotations, in turn, may be associated with either a "conservative" or a "liberal" leaning of the opinion.

The advantage of the whole opinion text as input is that no information is lost. Its downside, however, is that it may include more noise than only citations or quotations.

3.2.3. Text pre-processing

For any data subset, the raw text needs to be pre-processed. We applied the prevalent practice of removing capitalization, punctuation as well as stopwords. Furthermore, we reduced the words to their word stem, base or root form (stemming).

3.2.4. Feature engineering

The pre-processed text was tokenized, and the tokens were then used to form lists of n-grams (phrases) up to length three. N-grams extract information from text through local word order (Suen, 1979; Sidorov et al., 2014). In the next step, these tokens were mapped to a numerical representation. We computed counts and frequencies over n-grams. The second specification is to weight the counts (tf) by inverse document frequency (idf), which up-weights relatively rare words that could be more informative of topic or ideology.

Apart from converting opinion texts to vectors, we included the year the case was decided, the circuit at which the case was heard as well as the case type as assigned by the authors of the Songer database to the feature set, as well. Via grid-search, we established which input and pre-processing combinations worked best, especially regarding single words versus n-grams.

3.2.5. Model

After vectorization, the next step is the actual classification of the text input,[3] listed in the second last column of Fig. 2. In general, the classifiers may be grouped into two families, with the first being statistical methods. The advantages of this family are high explainability as well as being well-researched and understood (Ribeiro et al., 2016). The second family are deep learning algorithms mostly comprised of some form of neuronal network architecture. In common NLP tasks, these algorithms outperform traditional algorithms (Kim, 2014; Vaswani et al., 2017). However, a downside to these models is that feature introspection, as well as explainability, is difficult. While there are attempts to develop methods for feature introspection, such as Shrestha et al. (2017) or Ribeiro et al. (2016), results so far are preliminary. Consequently, we focus on well-researched statistical classifiers, maximizing the explainability of the results. The classifier, we deploy are a passive aggressive classifier (Crammer et al., 2006), a logistic regression (Schmidt et al., 2017), a ridge classifier (Rifkin and Lippert, 2007), as well as a support vector machine with stochastic gradient descent (SGD) learning (Zhang, 2004). All models are trained on a stratified train-test split with respect to case type.

3.2.6. Model evaluation

For model evaluation, we use standard performance metrics for machine learning, namely accuracy, precision, recall and f1-score (last column of Fig. 2).[4] The f1-score is the harmonic mean of precision and recall. As compared to accuracy for example, it is more stable with respect to unbalanced data-sets like ours. Furthermore, in the context of this paper we consider precision as more important as recall, because our dataset contains much less liberal than conservative cases. Thereby, we consider it as more important to actually find theses few liberals and risk to classify some conservatives as liberal.

As all performance measures are 5-fold-cross-validated, the scores reported are weighted averages. As the label space per category is heavily imbalanced in the validation set, accuracy has to be interpreted with care, and therefore the best performing classifier is selected by referring to the weighted f1-score. In our case, an additional model evaluation is the use of the predictions in the replication analysis below.

3.3. Evaluation of results

In the following, we provide in-depth analysis across the different classification models introduced by Fig. 2.

3.3.1. Performance metrics

Appendix C depicts the performance metrics f1-score, accuracy, precision and recall for all models tested. Fig. 10 shows that the scores depend more heavily on the subset-input-combination than on the specific classifier used.

Based on this observation, we select four models to analyze and compare in detail. Fig. 3 depicts the model for each of the four *subsets* tested which reaches the highest *f1-score*. We report the accuracy, precision, recall, and f1-score respectively (coded by color, see legend). Each of the four groups of bars refers to a different subset of the data, for which we explored different modeling approaches. The top row looks only at the liberal and conservative votes, dropping the "other" category. Second, we classify the full dataset with all three categories. Third, we limit the dataset to criminal cases. In the bottom row, we limit the dataset to economic cases.

On the *y* axis, we indicate a feature that all four models have in common: they perform best on the input *opiniontext*, rather than on citations or quotations. While additional calibration and tweaking of the model parameters would improve the performance of the classifier using either citations or quotations as input, the result is consistently outperformed when using the complete opiniontext as input. This observation contrasts with the idea that citations or quotations would summarize the information in a meaningful way. However, instead of subtracting what we considered as noise, it seems that these input variations subtract important information.

As mentioned, the four *subsets* differ with respect to the subset of cases. Comparing the subsets concerning *label*, we differentiate between two or three label classification. The subset displayed at the top of Fig. 3 takes two labels into account. A random guess, assuming a random distribution of labels, should yield an accuracy of approximately 1/2. The model reaches an accuracy of 67.04%, lying clearly above this threshold.

The second group of statistics are from the three-labels model. How much performance do we gain when predicting two instead of three labels? The two models at the top of Fig. 3 show – only these two take all case types into account – an increase in accuracy from 62.00% to 67.04%. We believe that this increase in performance may offset the loss of information by excluding the "mixed/other" label as less than 1/7 of all cases fall into this category. This opinion is shared by other authors, as well: Most studies drawing upon the Songer/Auburn do exclude the "mixed/other" cases. However, for the sake of thoroughness we undertake the calibration presented in the following section for both the two and three label subset.

In the third and fourth groups of performance metrics, we show the three-label model but subset on case type. Interestingly, performance depends strongly on the case type. As mentioned in Fig.

[3] The classifiers are implemented with the python package sci-kit learn and fall into the category of supervised learning.

[4] While in traditional statistics measures such as the p-value are more prevalent, that measure is not appropriate in machine learning because we are trying to form accurate test-set predictions rather than to test for treatment effects. Moreover, the features in machine learning are often very highly correlated, so the estimated coefficients for them are difficult to tease apart.

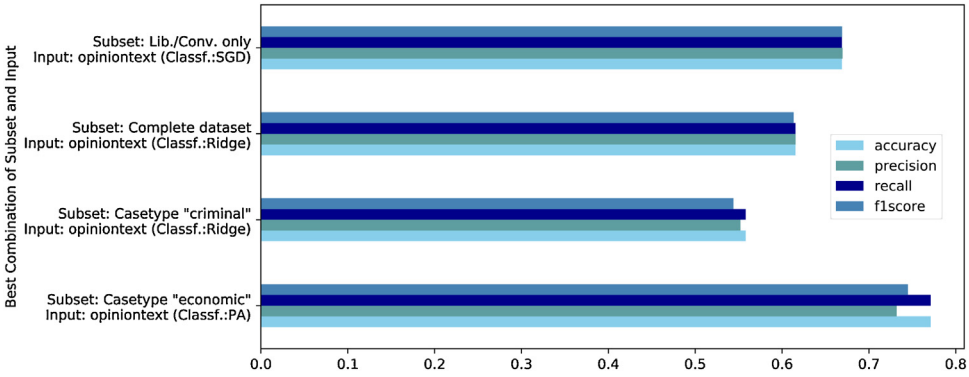


Fig. 3. Best performing combinations by subset. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

3.1, directionality is defined within case type while the number and quality of rules are quite distinct. Additionally, as Fig. 1 shows, case type is heavily imbalanced in favor of economic rather than criminal. These two facts help to explain why the subset *criminal* only reaches an accuracy of 55.80% and by contrast, why the subset *economic* achieves an accuracy of 77.10%. However, in order to increase generalizability, we instead opt to focus on classifiers trained on data containing all case types as some results from e.g. the case type "economic" may carry over to the case type "criminal".

3.3.2. Probability calibration

In the following, we analyze our classifiers' calibration: predicting a judicial opinion to either be conservative or liberal, we not only want to know the label but how confident the classifier is in assigning one particular label versus the other. In order to boost calibration, the classifiers were re-calibrated using either a sigmoid or an isotonic calibration function. The sigmoid function rests on a parametric approach based on Platt (1999)'s sigmoid model. The non-parametric isotonic variant is based on an isotonic regression.

Fig. 4 depicts the Ridge and SGD classifier respectively. For both classifiers, the calibration methods were applied for visualization purposes.[5] The three corners of Fig. 4 correspond to the three classes: conservative, liberal, and mixed/other. Arrows point from the probability vectors predicted by an uncalibrated classifier to the probability vectors predicted by the same classifier after calibration. For clarity of presentation, only each fiftieth data point from the test set is depicted.

Fig. 4 shows that calibration results in both classifiers shifting from under-confident to over-confident predictions. This can be seen as the mass of predicted points moves away from the center of $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ towards the edges. This means that the classifier is likely to categorize similar cases very differently as the predicted label is further away from the decision boundary for all cases. On the other hand, it also means that the classifier gets more confident about cases which are hard to classify – that is, the position of which is properly close to the decision boundary. We accept this change however, as the absolute accuracy as well as the f1-score increases, although there may be additional error for boundary cases.

While the two classifiers do not majorly differ in their confidence, they do differ in their error rate of assigning the label "liberal" to liberal cases. If one looks at the blue arrows, which depict cases for which the true label is "liberal", one can see that for the Ridge classifier (left panel) the mass of the blue arrows falls into the simplex spanned by the corner points $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$, (1, 0, 0). Every arrow point found within this simplex is classified as "liberal". Consequently, as the mass of blue arrows falls into that area, the mass of them is categorized correctly. In contrast, for the SGD classifier (right panel) a lower amount of the blue arrows falls into that area, meaning that the misclassification rate for "liberal" is higher. This means the precision for liberal is lower for SGD compared to the Ridge classifier. On the other hand, the inverse is true for the recall. As the original dataset features fewer liberal cases than conservative, on balance we might prefer to mis-classify conservative cases as liberal instead of liberal ones as conservative. At this point, this speaks in favour of the Ridge classifier versus the SGD classifier.

When looking at the "mixed/other" cases, we can see that the Ridge classifier classifies the majority of them correctly. However, that seems to come at the expense of mis-classifying a disproportionally high amount of liberal cases. For the reasons stated above, we consequently exclude the "mixed/other" label to gain performance in predicting only the labels "conservative" and "liberal".

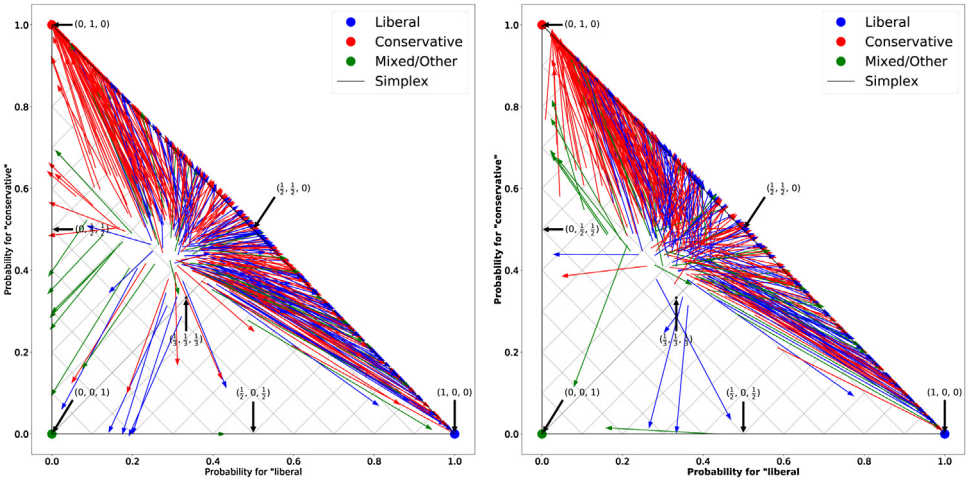
Fig. 5a provides another visualization to assess how well the probabilistic predictions of different classifiers are calibrated: it displays reliability curves which show the correct proportion of conservative cases (vertical axis) against the bins of predicted probabilities that a case is conservative (horizontal axis). The closer the reliability curve is to the 45-degree line, the better is the classification model's performance in terms of reproducing the original distribution. The Ridge classifier with isotonic calibration, as well as the SGD classifier with sigmoid calibration are highlighted in shades of blue.

Consider the Ridge classifier: For all cases which it predicts to be conservative with a 20% probability, about 40% are actually conservative. In other words, it underestimates conservativeness. However, for cases close to the hyperplane (0.5 probability for either directionality), Ridge approximates the directionality distribution very well.[6] Finally, at around 70% likelihood, the classifier begins to overestimate the number of conservative cases.

Alongside Fig. 5a, Fig. 5b shows that despite calibrating the classifiers, a significant part of the predicted directionality's mass lies close to the decision boundary of 0.5. This, in turn, means that the classifiers have to be relatively precise close to the decision boundary and be able to shift away mass from the decision boundary. Fig. 5b shows that the two classifiers most successful in this are the

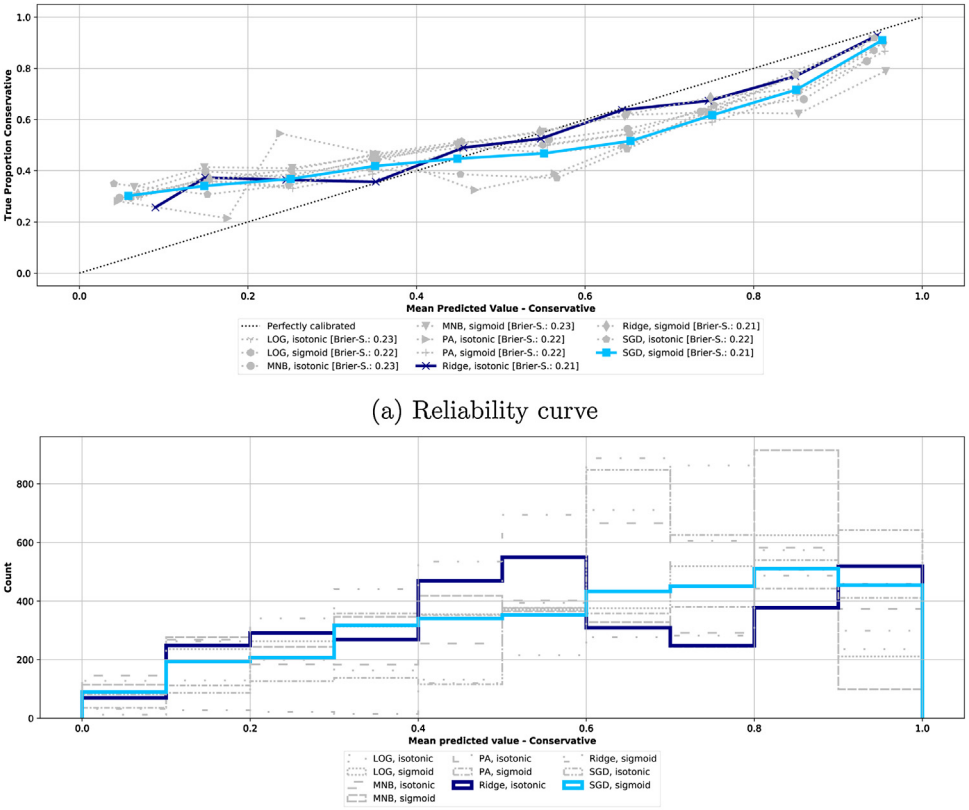
[5] Probability calibration was performed on data not used for model fitting. To this end, the training set consisting of 80% of the Songer data was cut in thirds and the model was then trained with 3-fold cross-validation. During this, 2/3 of the data were used for training and 1/3 was used for calibration. For each classifier the calibration algorithm yielding the best results was chosen.

[6] This is an important aspect as the Ridge classifier is similar to a support vector machine in that it uses the instances closest to the hyperplane for the separation of the data points.



(a) Ridge (Isotonic Calibration) (b) SGD (Sigmoid Calibration)

Fig. 4. Drift-plots showing the change of predicted probabilities after calibration. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)



(a) Reliability curve

(b) Distribution Diagram

Fig. 5. Reliability curves and distribution diagram. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

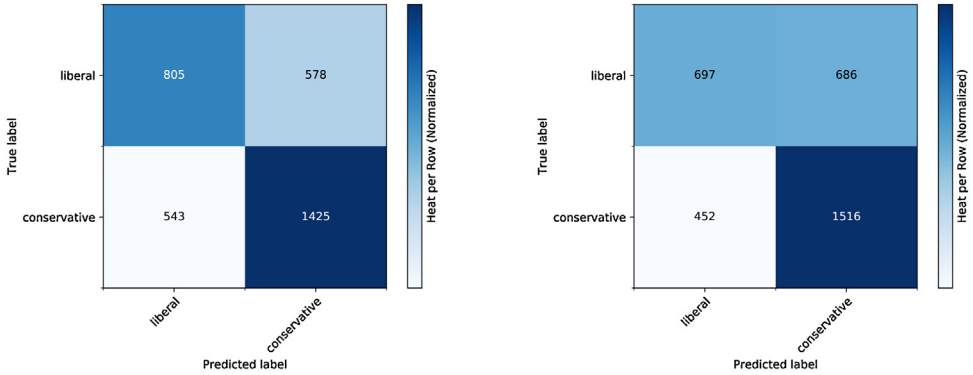
ridge classifier, calibrated with the isotonic algorithm, and the SGD support vector machine, calibrated with the sigmoid algorithm.

3.3.3. *Heatmaps*

In the previous paragraph, we conclude that a two label classifier for all case types will be the basis for predicting political ideology labels. In terms of performance metrics, the SGD classifier reaches the highest f1-score. However, the decision for the final model should not just take the f1-score but rather the types of errors that the classifier makes into account, as well. Therefore, Fig. 6 plots normalized[7] confusion matrices for those two models deploying the best f1-score: The Ridge as well as the SGD classifier.

As mentioned in Section 3.2, we consider it as crucial to correctly predict as many liberal cases as possible, even if some conservative cases are wrongly predicted as liberal. Fig. 6b shows that as

[7] Normalized heat is calculated by dividing each value by the row mean.



(a) Ridge classifier (b) SGD classifier

Fig. 6. Confusion matrices for the classifiers SGD and ridge.

far as liberal cases are concerned, the SGD classifier predicts 697 cases correctly as liberal but almost as many cases (686) wrongly as conservative. The Ridge classifier displayed by Fig. 6a, by contrast, predicts 805 liberal cases correctly as liberal and only 578 liberal cases wrongly as conservative.

3.3.4. *Best classifier*

Based on performance metrics, heat-maps, and calibration results, we can select the classifier most suited for the task set out in this paper. The f1-score – our preferred performance metric – peaks for the Ridge-Classifier, calibrated with an isotonic function as well as for the SGD-classifier, calibrated with a sigmoid function. The second performance metric we consider as critical is precision, for which the Ridge classifier shows better results than SGD. In the same vein, the reliability curves show that Ridge is closer to the 45-degree line than SGD, which makes the former preferable. The only aspect where the SGD support vector machine slightly outperforms the Ridge classifier is in terms of mass, as shown in Fig. 5b. However, overall, the difference in this regard is negligible. Given this reasoning, we chose the Ridge-classifier calibrated with the isotonic algorithm as model to perform out of sample predictions.[8]

3.4. *Analysis*

This section analyzes and interprets the predictions of the best two-label classifier. We look at predictions over time and by judge. We also interpret the model by examining predictive features.

3.4.1. *Prediction of the time series in decision direction*

Landes and Posner (2009) point out that the accuracy of the original Songer data is susceptible to the year in which a judge decided a case. Coders had more trouble coding older cases as compared to newer ones. We would like to see if this is reflected in differential performance of our classifier over time.

Fig. 7 shows the fraction of conservative and liberal cases by year for all circuits.[9] We include out-of-sample data which is made up of scraped lexis data without the cases already within the Songer data set. The original scraped data set holds more than 1 million cases. As our classifier uses the year of the case, the circuit, and the case

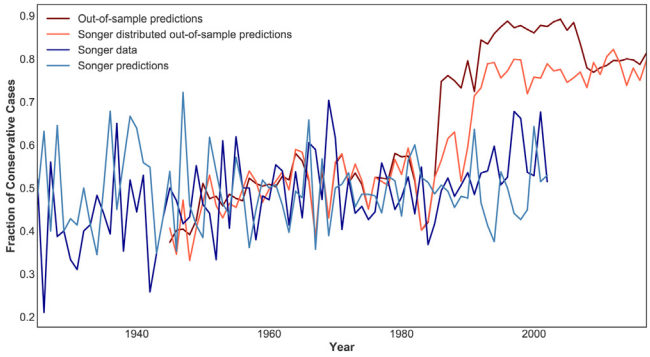


Fig. 7. Fraction of conservative and liberal cases, each calculated for actual as well as predicted case directionality, plotted by year. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

type as laid out by Songer[10] these features have to be available for all out-of-sample cases, as well. Especially the last one constrains the lexis data set because case type was only available for cases of the years 1930 and later. Consequently, Fig. 7 shows out-of-sample predictions only for those years.[11]

Fig. 7 shows that for the in-sample predictions on the test set of the Songer data (20% hold-out data), the predictions closely approximate the original labels. This is also reflected in the high correlation of 0.73 ($\alpha < 1\%$). Especially for the years 1950 to 1980, the classifier performs very well. The out-of-sample predictions for that time period approximate the trend observed in the Songer data. Only for the years of 1980 on-wards, the out-of-sample data (red line) is predicted to be considerably more conservative.

This spread may be caused amongst others by the classification error. Another reason could be the sampling process used by Songer and his team to construct the database.[12] To test this presumption, we plot a subset of the lexis data constructed according to Songer's rules ("Songer-distributed out-of-sample", the orange line). Indeed,

[8] The final specifications of the classifier are as follows: We preprocess the text by excluding all stop words as well as punctuation. Following that, a lemmatizer is applied. This input transformed into bigrams and then fed to a tfi-vectorizer. That vectorizer calculates the distance based on the "l2"-norm. It also makes uses the three additional features of year, circuit and case type. The regularization strength parameter α for the Ridge classifier is 2.0.

[9] The cases categorized as "mixed" or "other" are excluded.

[10] We matched the lexis case types to the one laid out in the Songer database. However, the match has no bijective property. In order to get a reasonable good match, the subcategory case types of both, the Lexis database as well as the Songer database were used. This match is surjective with the Lexis subcategory case types as a base set. Then the matched Songer sub categories are aggregated to a Songer top category. Except for very few cases (<1000) this aggregation is unequivocal.

[11] If one is willing to forgo the performance gain introduced by the case type feature (about 2.5% points in the current configuration), one can predict directionality for all lexis cases.

[12] For the original Songer database, at maximum 30 cases per year per circuit were sampled from all available cases after 1961. Before 1961, only 15 cases per year per circuit were selected.

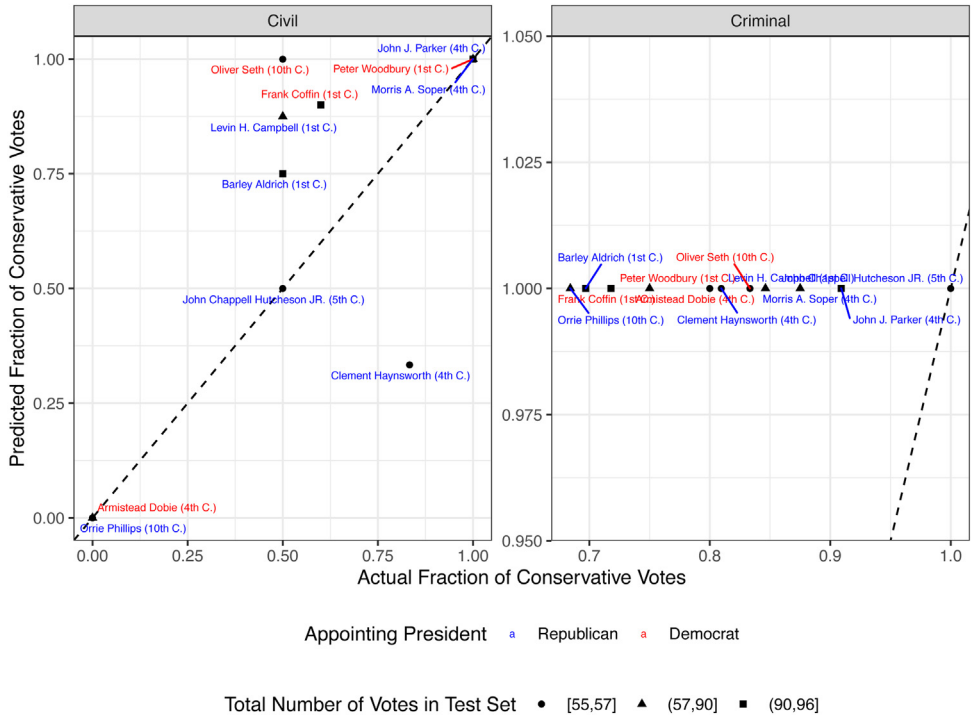


Fig. 8. Fraction of directed votes per judge – comparison actual votes and predicted votes.

we find that the orange and red lines diverge after 1980, with the orange line being closer to the original Songer data. This illustrates that indeed the sampling process heavily influences the distribution of decision directionality: As soon as the total amount of cases increases[13] by a significant amount, a spread appears. As the absolute number of court cases increased over time (Casper and Posner, 1974), at least for cases after 1980 the Songer data may not be a good sample for the full set of cases. Consequently, the difference in out-of-sample predictions as compared to Songer predictions may simply stem from the fact that there is a structural shift in conservativeness (either in variation or trend) from 1980 onwards which is not represented by the Songer sample.

3.4.2. Directed votes per judge

Next we zoom in on particular judges. We look at performance for the ten judges who cast most of the votes in the Songer data set, analyzing performance in civil and criminal cases separately. Those judges who did not hear both civil and criminal cases were excluded. The horizontal axis of Fig. 8 indicates the true proportion of conservative votes, while the vertical axis indicates the predicted proportion of conservative votes. Each point indicates these statistics for a single judge. If a judge's predicted behaviour is the same as the truth, then his/her data point would lie on the dotted 45-degree line.

Fig. 8 shows that for civil cases, predicted and actual fractions are quite close. A χ^2 test shows that the distribution of predicted fractions is not statistically different from the distribution of actual fractions ($p^{\chi^2} > 0.1$). For case type criminal, however, the distributions of true and predicted fractions across judges are statistically different. The reason for this might be that the majority of criminal cases is labeled as conservative. Consequently, as the classifier uses the case type as feature it can increase performance on criminal cases by labeling it as conservative. In other words, the classifier tends to overpredict the number of conservative cases in criminal law.

3.4.3. Feature inspection

To further understand the two-label classifier, we investigate the features that are most important in driving our predictions.

For this purpose, let *feature* be a feature, *value* be a value it could take, and *label* one of the ideological directions (conservative or liberal). We ranked the informativeness of each feature by the highest value of $P(feature = value|label = conservative)$ divided by $P(feature = value|label = liberal)$. Note that these are equivalent to coefficients from a Naive Bayes Classifier. The coefficients of the different features are represented by their standardized moments, meaning that normalization was performed by dividing through the standard deviation. This means that each coefficient is on the same scale and therefore comparable. The hyperplane separating "conservative" from "liberal" lies at 0, meaning a hypothetical case for which all the decision results would be zero falls into neither category. The higher the coefficient of a feature, the further away does a single feature move the case instance from the hyperplane when the feature is present within the case.

Table 1 lists the most informative features used by our best performing classifier. Please note that the *most informative* features for the label "liberal" are constructed such that they are *least informative* for the label "conservative". The features are either opinion-text phrases, quotation phrases, or citations.

Table 1 shows that the coefficients differ vastly in absolute size across the three different input variations. This corroborates the results of the metric scores. Especially for citation as input, the range of the coefficients' values is vary narrow, with −7.49 being the minimum and 10.16 being the maximum. Consequently, many features loading clearly either the "liberal" or the "conservative" side are needed in order to have the case fall into a category. By contrast, the range of the coefficients' values for opiniontext is much wider, with a minimum of −57.67 and a maximum of 189.96. A case including the words "reverse remand" for example would be classi-

[13] Where for the year 1945 only slightly more than 100 cases per year per circuit were coded with a usable case type in the out-of-sample dataset, for the year 2000 there are more than 2000 per year per circuit.

Table 1
Best predictive features.

	Quotations (Ridge)		Citation (Ridge)		Opiniontext (Ridge)	
	Coef	Feature	Coef	Feature	Coef	Feature
Best predictive features for label "Conservative"						
1	−17.13	knowingly	−7.49	Humphrey_v_Moore	−57.67	motion new
2	−13.18	John_Doe	−7.43	Dandridge_v_Williams	−53.71	plaintiff argue
3	−11.97	unique_circumstances	−6.59	SEC_v_Chenery_Corp	−51.91	prior art
4	−11.47	X	−6.42	Co_v_Zenith_Radio_Corp	−50.86	appellant claim
5	−11.40	No	−6.19	Dalehite_v_United_States	−50.78	grant motion
6	−11.03	minor	−6.06	Brady_v_Maryland	−49.45	plaintiff appellant
7	−10.85	search	−5.60	United_States_v_Robinson	−48.85	plaintiff contend
8	−10.63	attractive_nuisance	−5.55	Mal_v_Riddell	−45.70	fiduciary duty
9	−10.09	may	−5.38	Port_Gardner_Investment_Co_v_U	−45.62	plaintiff appeal
10	−10.04	overhead	−5.25	Olim_v_Wakinekona	−44.01	judgment affirm
Best predictive features for label "Liberal"						
1	19.98	that_where_the_State_has_provided_an_opportuni...	10.16	Yes_v_United_States	189.96	reverse remand
2	19.86	Motion_for_Judgment	9.18	United_States_v_Taylor	133.90	remand proceeding
3	19.57	fairer_to_those_adversely_affected_by_a_bond_f...	9.11	...Inc_v_Commissioner	103.28	case remand
4	19.16	take_care	9.09	Townsend_v_Sain	98.70	remand district
5	18.30	urge_that_the_indictment_charged_the_maintenan...	8.88	United_States_v_Young	89.69	government argue
6	17.32	good_faith	8.43	Dennis_v_United_States	85.99	remand new
7	17.30	anything_of_value	8.21	Coppedge_v_United_States	84.05	proceeding consistent
8	16.76	crack a_little_bit_of_time_to_research_on_the_...	8.15	...Inc._v_United_States	75.33	consiStatent opinion
9	16.76	a_little_bit_of_time_to_research_on_the_backgr...	8.00	Green_v_United_States	74.29	new trial
10	15.49	clear_and_convincing	7.97	Brown_v_Board	60.13	reverse case

fied immediately as liberal. In essence, this means that features for the opiniontext or quotations as input are more informative than for the citations.

The first column of Table 1a and b have the most predictive quotations. Quotations loading heavily on the label "conservative" are "knowingly" or "unique circumstances". The court quotes these phrases, i.e. they are singled out as relevant to the case at hand. Both phrases indicate a possible conviction. As the code book by the authors of the Songer database very often label a conviction as "conservative", this seems to be in line with the data provided. On the other side, the quotations for "liberal" are not as easily interpreted.

The second column of Table 1a displays those citations loading on the label "conservative". For the most heavily conservative citation, Humphrey v. Moore, the court limited the power of unions from infringing too far on employees of a company not part of the union. In Dandridge v. Williams, the court found that the state has some right to interpret how it puts into practice federal welfare laws. In consequence, Maryland was found not to be in violation of the anti-discrimination act. Another conservative example would be United States v. Robinson, in which the court strengthened the police powers for searches during lawful arrests under the fourth amendment.

In comparison, in the second column of Table 1b features citations which the classifier finds to be indicative of a liberal case. The most indicative citation would be United States v. Taylor, a case in which the bar for conviction on charges of conspiracy was raised. Coppedge v. United States dealt with the fact that the sentenced petitioner had not received the plenary review of his conviction to which he is entitled and all his appeals against his conviction against this ground were dismissed. The Supreme Court reversed the decision to dismiss his appeal and generally strengthened defendants rights in this regard. In the same vein, Green v. United States reversed the sentencing of the defendant under the Fifth Amendment as he was put in jeopardy twice for the same offense. Consequently, while absolute size of the coefficients for citations hint at only a limited quality for the overall classification into either "liberal" or "conservative", the cases as such seem to fall into the right domain.

The last column shows the predictive phrases from the full opinion text. Features such as "judgment affirm" or "plaintiff appeal" are predictive of the conservative label. In line with those but not shown here are the features "affirm judgment" and "appeal dismiss" on place 11 and 14 respectively. This is in line with labelling rules as set out by the Songer team for criminal cases, where the coding rules state that affirming the decision against an appellant is to be coded as conservative. Conversely, within the most predictive features for "liberal" one can find "reverse remand", "remand proceeding", or "reverse cased", reflecting that predictive features seem to be driven by criminal cases.

4. Replication and robustness checks

This section focuses on the replication aspect of Landes and Posner (2009). For comparison, all tables and figures that Landes and Posner (2009) produce with data of Circuit Courts are listed in Table 5, Appendix A. The most relevant tables for our purposes are Tables 11 and 13 as numbered in the original paper.

Summary statistics. This paragraph compares our summary statistics listed in Table 2b to those by Landes and Posner (2009, p. 803) listed in Table 2a. As can be seen, the statistics differ. We count a total of 56,602 cases; Landes and Posner (2009) count 55,041 cases. Furthermore, we count more opinions classified as "conservative" or "other" than Landes and Posner (2009) do.

One possible explanation for these diverging results is that not all of corrections that Landes and Posner (2009) applied in the original paper were described in sufficient detail to reproduce. We were able to apply the corrections concerning political ideology (Landes and Posner, 2009, pp. 830–831) but we were unable to apply judge-related corrections. Landes and Posner (2009) briefly mention judge-related corrections and refer to a website for a detailed description. This website however, is no longer available online.

Regression. Next we replicate the primary regression analysis of circuit court judges in Landes and Posner (2009), focusing only on the essential part of their analysis. For Table 13, we replicate the

Table 2
Court of appeals votes by subject matter and ideology for 538 court of appeals judges only: 1925–2002.

	Crim	Civ Rts	First	Due Proc	Priv	Labor	Econ	Misc	Total
Original by Landes and Posner (2009)									
Conservative	6823	2721	566	461	117	1351	9361	525	21,925
Liberal	1876	1766	477	201	67	1922	9884	559	16,752
Mixed	635	460	89	51	13	420	1775	22	3465
Other	5321	210	102	79	3	179	6047	958	12,899
Total	14,655	5157	1234	792	200	3872	27,067	2064	55,041
Replication									
Conservative	7217	2647	397	412	83	1397	11,084	478	23,715
Liberal	1911	1755	379	176	38	0	10,375	596	15,230
Mixed	613	473	86	48	9	423	1689	31	3372
Other	5652	212	40	24	3	2232	5177	945	14,285
Total	15,393	5087	902	660	133	4052	28,325	2050	56,602

regressions focusing on the fraction of conservative votes and only taking the period from 1925 to 2002 into account.[14]

Regarding the baseline regression, Landes and Posner (2009) specify their regression model as follows:

$$FrCon_{ij} = \beta_0 + \beta_1 X_i + w \quad (1)$$

where $FrCon_{ij}$ denotes the fraction of conservative votes, calculated as votes per judge over the sample period. X_i encompasses several judge characteristics such as the party of the appointing president, share of Republican senators at the time of nomination, year of appointment, gender, race,[15] prior experience as a district judge, as well as judge circuit fixed effects[16] According to Landes and Posner (2009, p. 810), their regressions are weighted either by the judge's total votes in civil cases or the total votes in criminal cases. Furthermore, Landes and Posner (2009) do not specify how they compute their standard errors, but we assume that they use heteroskedasticity-robust standard errors (treating each judge as an observation) and therefore use errors of that type for the replication.

4.1. Civil cases

In Table 3 we provide our first replication table, dealing with civil cases only. Column (1) corresponds to Landes and Posner (2009) Table 13 column (6).[17] As in the original paper, we report the *t*-statistics, rather than standard errors or *p*-values, for all coefficients in parentheses. Landes and Posner (2009) do not specify how they computed standard errors for their regression Table 13, but we inferred that they used heteroskedasticity-robust errors.

The main research interest of Landes and Posner (2009) was whether judges follow their party affiliation in their decisions. They find a significant influence of being appointed by a Republican president *(RepPres)* on the fraction of conservative votes for civil cases (Table 3, column 1). Our result for civil cases (Table 3, column 2), is quite similar when compared to Landes and Posner's; in our data, being appointed by a Republican is associated with a positive and significant effect of voting conservatively in civil cases. The evidence for a relationship between party and ideology actually appears to be stronger in our replication than implied by the original study.

Apart from deploying heteroskedasticity-robust errors, we propose a model specification with multi-way clustering (non-nested) as recommended by Cameron et al. (2006). Based on the advice from Abadie (2018), we add two-way clustering by circuit and year. This allows for correlation in the error term across judges within court over time, as well as across courts in the same year. Clustering leaves coefficients unchanged, and a comparison of columns (2) and (3) reveals that *t*-statistics only differ slightly as a result of the two-way clustering.[18]

While Landes and Posner (2009) grouped the data on judge-level, we additionally run the empirical analysis with data at the vote level. This specification allows us to control for case characteristics with circuit-year fixed effects. For getting at the effect of party affiliation on ideology, this is an important step econometrically because the number of Republican-appointed judges and the proportion of conservatively decided cases could be correlated over time due to unobserved confounding factors.

The dependent variable is now binary. It equals one for conservative decisions and zero for liberal decisions (cases with the mixed/undetermined category are dropped). The vote level regression model includes circuit-year fixed effects, as well as clustered standard errors by judge and year.

This specification successfully replicates the significant positive effect of a conservative appointing president (*RepPres*) on the fraction of conservative votes.

Model specifications (5) and (6) are estimated not only with hand-labeled but also with predicted data. The predictions on which estimation results of columns (5) and (6) are based, were generated with a calibrated Ridge classifier.

These re-estimations serve as an alternative way to assess the performance of the classifier. The rationale behind this procedure is that generating labels is not the end-goal, but using these labels in an empirical model is. Therefore, even if the classifier cannot predict political ideology with an accuracy of 100 percent, its performance can be viewed as appropriate if the results of the empirical model do not change drastically when estimated with the classifier's predictions.

As far as column (5) is concerned, using predicted instead of hand-labeled data does not change the results for coefficients *RepPres*. Estimating the vote level fixed effects model with predicted labels instead of hand-labeled (column 6) results in estimates for RepPres that are no longer statistically significant.

[14] In turn, this means that we do not display results for the fraction of liberal votes, as displayed in columns (2) and (4) of Landes and Posner (2009) Table 13, nor do we report results for the period of 1960 to 2002 as reported in Table 14.

[15] Race is a dummy for Black = 1, 0 else.

[16] The judge specific data was acquired from the Auburn database by Gary Zuk, Deborah J. Barrow and Gerard Gryski on songerproject.org/attributes.html and then matched to the Songer data by a judge identifier code.

[17] These are the columns with the "uncorrected" data. We only compare uncorrected data as Table 2 showed that we were not able to replicate even summary statistics for the corrected version.

[18] We provide regression results with errors clustered on the year of appointment, Circuit Court, and the party of appointing president in Table 3, column (3).

Table 3
Regression analysis of court of appeals votes: 1925–2002, civil cases.

	Dep. variable: fraction of conservative votes					
	true data				*predicted data*	
	Landes (2009) (1)	Replicated (2)	Clustered (3)	Vote (4)	Clustered (5)	Vote (6)
RepPres	0.035***	0.069*	0.069***	0.092***	0.032**	0.031
	(3.860)	(2.125)	(4.136)	(3.821)	(2.942)	(1.417)
SenRep	0.072	−0.017	−0.017	0.095	0.004	0.219
	(1.710)	(−0.090)	(−0.347)	(0.647)		(1.677)
YrAppt	0.0003	0.001	0.001	0.0003	0.001	0.001
	(0.790)	(0.665)	(1.237)	(0.202)	(0.796)	(0.431)
Gender	−0.006	0.015	0.015	−0.026	−0.0004	−0.058
	(0.260)	(0.344)	(0.318)	(−0.681)	(−0.011)	(−1.384)
Black	−0.028	−0.105	−0.105	0.007	−0.125	−0.001
	(1.180)	(−1.505)		(0.124)		(−0.023)
DistrictCourt	0.002	−0.004	−0.004	−0.002	−0.002	−0.0005
	(0.330)	(−1.455)	(−1.183)	(−1.712)	(−0.345)	(−0.417)
FracEcon	−0.090	−0.230	−0.230**	0.355**	−0.249	0.451***
	(1.640)	(−1.506)	(−2.690)	(2.774)	(−1.918)	(3.531)
FracMisc	−0.049	1.345*	1.345*	−0.920	1.464***	−0.324
	(0.350)	(2.442)	(2.107)	(−1.842)	(6.118)	(−0.673)
Circuit FE	Yes	Yes	No	No	No	No
Circuit-year FE	No	No	Yes	Yes	Yes	Yes
Observations	535	498	498	4169	498	4169
R^2	0.240	0.119	0.119	0.047	0.123	0.066

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.
Linear regression with heteroscedasticity robust standard errors.
Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge; *DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: all regressions include 11 dummy circuit variables – circuits 1 to 11 with the D.C. court the omitted circuit variable.

4.2. Criminal cases

In Table 4, we provide our second replication table; it deals with criminal cases only. Landes and Posner (2009) found a positive and significant influence of being appointed by a Republican president *(RepPres)* on the fraction of conservative votes. Our result for criminal cases is quite similar to Landes and Posner's, our coefficient being slightly larger. Furthermore, for criminal cases, Landes-Posner found a negative effect of appointment year. However, we do not find such an effect. They also report a negative impact of being black (*Black*) on crime conservatism, which we replicate.

Two-way clustering changes *t*-statistics only slightly. This leads to no change in significance level for the coefficient *(RepPres)*, but it left the coefficient *(Black)* to no longer be significant.

The fixed effects multi-way clustering model on vote level data replicates the significant and positive effect of the party of the appointing president (*RepPres*) as well as of being black (*Black*) on the fraction of conservative votes.

The multi-way error component model using predicted data could not reproduce the significance of the coefficient *RepPres*. Instead, being male turned to have a significant negative impact on criminal conservatism. The fixed effects multi-way clustering model on vote level with predicted data could neither reproduce the significance for coefficient *RepPres* nor *Black*.

4.3. Extreme bounds analysis

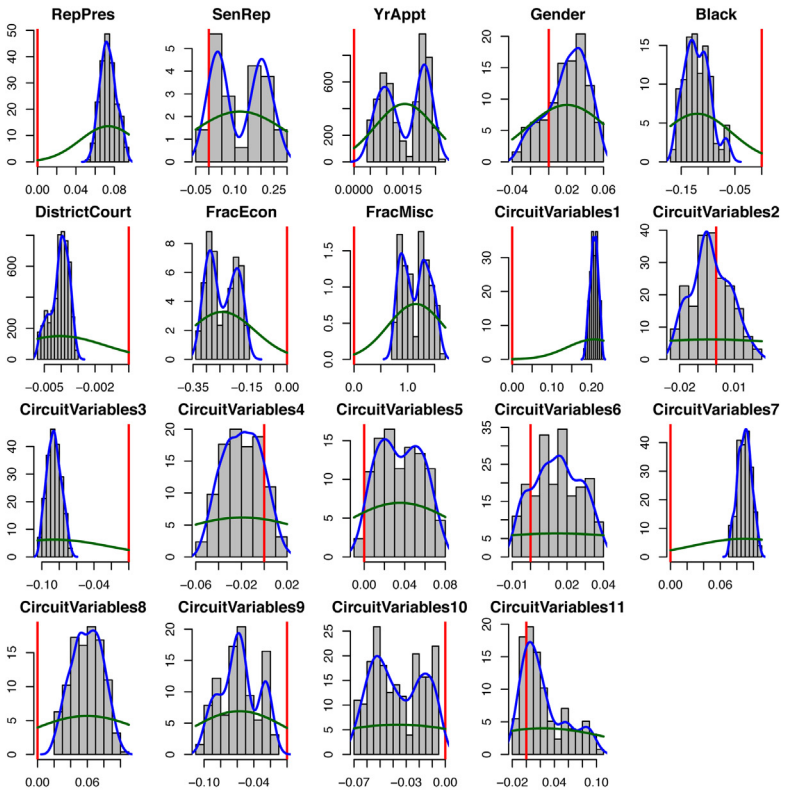
The extreme bounds analysis (EBA) is a sensitivity test that examines how robustly the dependent variable of a regression model is associated with a variety of possible determinants (Hlavac, 2016). We estimate an EBA, including all possible combinations of independent variables that Landes and Posner (2009) specified. To limit the influence of coefficient estimates with high multicollinearity, we follow the recommendations by Hlavac (2016) and specify the maximum acceptable variance inflation factor to be 7. Next, we increase the weights of those regression models that better fit the data – that is, by its likelihood ratio index according to McFadden (1973).

Fig. 9 shows histograms for each of the independent variables included in the model. The green curve displayed in each histogram is a density curve which approximates the coefficients' distribution with a normal distribution.

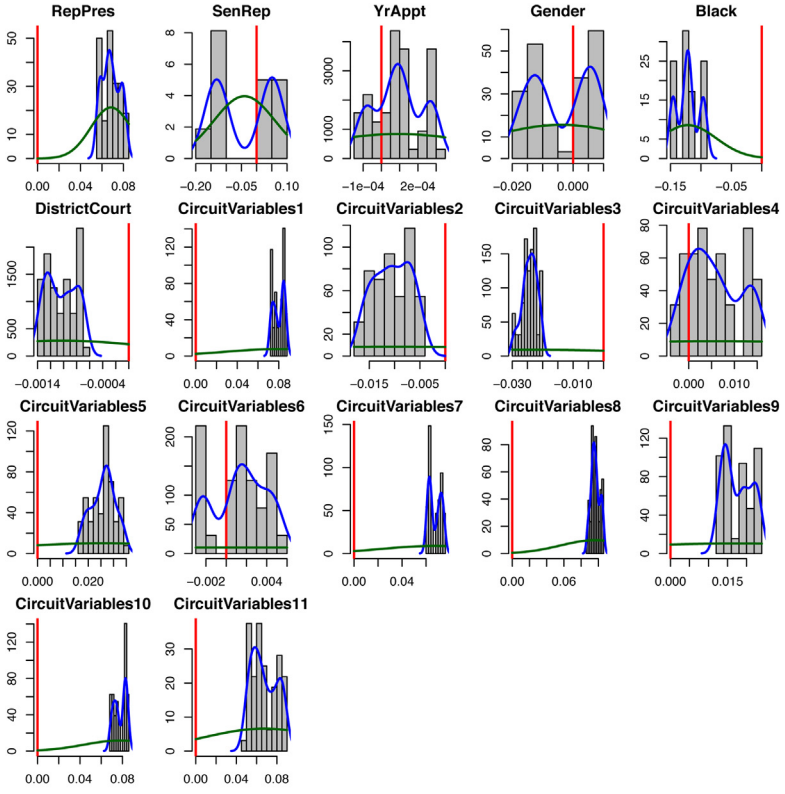
A positive coefficient indicates that holding all else equal, a higher value of the examined variable is associated with a higher fraction of conservative votes. On the other hand, if most of the area of the histogram's bins lies to the left of zero, higher values of the corresponding variable are associated with a lower fraction of conservative votes. For the civil cases, Fig. 9a suggests that when the appointing president *(RepPres)* is Republican (rather than Democrat), when the judge was appointed in later years *(YrAppt)*, as well as when the specific judge participated in a higher fraction of miscellaneous votes *(FracMisc)*, a judge's fraction of conservative votes increases. Furthermore, circuits 1 and 7 are consistently associated with a higher fraction of conservative votes. Being black *(Black)*, having served more years as a district judge *(DistrictCourt)*, as well as an increasing fraction of economic votes *(FracEcon)*, are associated with a lower fraction of conservative votes. Furthermore, circuits 3, 9, and 10 have a lower fraction of conservative votes.

To conclude the visual inspection as well as the interpretation of the statistics, found in Appendix E, the EBA for civil cases suggests that the variables *RepPres*, *FracMisc* and *circuit 1* are very strongly associated with the dependent variable.

For criminal cases, Fig. 9b shows that being appointed by a Republican (rather than Democrat) president (*RepPres*) is consistently associated with a higher fraction of conservative votes for all regression models estimated. Furthermore, circuits 1, 5, 7, 8, 9, 10, and 11 are associated with a higher fraction of conservative votes. By contrast, being black (*Black*) as well as having served more years



(a) civil cases



(b) criminal cases

Fig. 9. Histograms extreme bounds analysis, for civil and criminal cases. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

Table 4
Regression analysis of court of appeals votes: 1925–2002, criminal cases.

	Dep. variable: fraction of conservative votes					
	true data				*predicted data*	
	Landes (2009) (1)	Replicated (2)	Clustered (3)	Vote (4)	Clustered (5)	Vote (6)
RepPres	0.056**	0.077***	0.077***	0.051**	0.038	0.005
	(4.220)	(3.634)	(3.811)	(3.022)	(1.734)	(0.829)
SenRep	−0.076	−0.151	−0.151	0.010	−0.020	0.078**
	(1.090)	(−1.399)		(0.141)	(−0.542)	(2.844)
YrAppt	−0.001***	−0.00001	−0.00001	−0.0003	0.001**	−0.001**
	(3.390)	(−0.023)	(−0.032)	(−0.601)	(2.876)	(−2.709)
Gender	−0.014	−0.019	−0.019	0.010	−0.023*	−0.012
	(0.710)	(−0.740)	(−0.876)	(0.545)	(−2.219)	(−1.750)
Black	−0.057*	−0.091*	−0.091	−0.081**	−0.020	−0.027
	(2.060)	(−1.814)	(−1.047)	(−2.717)	(−0.257)	(−1.697)
DistrictCourt	0.001	−0.001	−0.001	0.0003	−0.001	0.001
	(0.140)	(−0.817)	(−0.390)	(0.360)	(−0.346)	(1.917)
Circuit FE	Yes	Yes	No	No	No	No
Circuit-year FE	No	No	Yes	Yes	Yes	Yes
Observations	523	498	498	13,543	498	13,543
R^2	0.240	0.084	0.084	0.019	0.052	0.014

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.
Linear regression with heteroscedasticity robust standard errors.
Variables: *RepPres*: Party of the appointing president, conservative or liberal (omitted category); *SenRep*: Share of republican senators at the point of election; *Gender*: sex of the judge, male or female (omitted category). *Black*: dummy for the race of the judge;*DistrictCourt*: Years spent as a district judge; *FracEcon*: Fraction of economic votes; *FracMisc*: Fraction of miscellaneous votes; *Circuit Variables*: all regressions include 11 dummy circuit variables – circuits 1 to 11 with the D.C. court the omitted circuit variable.

as a district court judge (*DistrictCourt*) decrease the fraction of conservative votes. Furthermore, circuits 2 and 3 are associated with a lower fraction of conservative votes.

To conclude the visual inspection, as well as the interpretation of the statistics found in Appendix E, EBA results for criminal cases suggest that the variables *Pres*, *Black* as well as *circuit 8* and *10* are robustly associated with the fraction of conservative votes.

5. Conclusion and outlook

This paper has two main goals. Our first goal was to replicate the analysis on Circuit Courts proposed by Landes and Posner (2009), and to add multiple robustness checks to assess the validity of the regression model initially specified. Second, we show an approach for extending the data set used in the original study via machine learning, especially in regards to the input used for any future algorithm.

As far as replication of the empirical analysis of Landes and Posner (2009) is concerned, we were able to reproduce the most critical findings. The robustness checks found, just as Landes and Posner (2009) did, that the party of the appointing president and being black influences the fraction of conservative votes. We find that the result for party affiliation is actually stronger than the original article found, as it extend to both civil and criminal cases.

What explains our different results? We paid particular attention to the code generating the fraction of conservative votes. As multiple reshaping and grouping operations as well as joining different datasets were necessary in order to obtain this variable, its calculation is not exactly trivial. We can imagine that a small mistake in the original code by Landes and Posner (2009), such as an inner instead of an outer join, could change the fraction. In turn, its association with the dependent variable may also change.

However, we could not replicate the exact summary statistics of the data set Landes and Posner (2009) used because they did not provide replication code and did not sufficiently specify their corrections in the original paper. That, in particular, may affect the rest of their findings.

In order to extend the data set, we experimented with different classifying algorithms, where the best one was a passive-aggressive classifier for economic cases, reaching an f1-score of 74.49%.

In order to assess the validity of the classification, we compared the regression results obtained by using predicted data to those obtained by using only hand-labeled data. Coefficients found to be significant with the replication as well as with the robustness checks were not replicated with the predicted data, suggesting that that (1) the classifier still needs improvement, or (2) researchers should be careful with using predictions as data in downstream empirical analysis. Future research should, therefore, take into account that the distribution of the Songer data in regards to cases per circuit per year does not mirror the distribution of the universe, and as such it may skew the predictions of any classifier. Oversampling is only an imperfect correction for this issue, as is the inclusion of the circuit or year as a feature. Otherwise, the consistency of results may not be guaranteed.

One aspect that we neglected thus far is that predictions cannot be directly plugged into a regression without correcting for the classification error. Fong and Tyler (2018) proposed one approach to do so. However, Fong and Tyler (2018) describe a case in which one or more independent variables are predicted. In our case, however, we predict the dependent variable. Therefore, we propose to develop a correction approach in order to prevent forward propagation of the prediction error used within a dependent variable which at this point may be of the main reason for failure.

Furthermore, the distributions of the enlarged data set and that one of the original data are significantly distinct. Overall, the classifier was trained on roughly 0.5 percent as compared to the number of labels that were predicted. As soon as such a considerable dissemblance is present, non-random draws or the lack of stratification is very problematic. Lack of stratification is the case with the original Songer database, i.e. Songer (1993) does not keep the original distribution of cases per circuit as they focused on preserving other aspects such as the presence of all circuits in each year.

Taking the above into account, our results provide a concise groundwork for future research in this area. First, in order to estab-

lish a ground truth that goes beyond mere statistical significance and also looks at distributional aspects more than just regression results are needed. Here, we suggest that taken our results multiway error component modeling as well as an extreme bounds analysis should be used on any prior results before trying to take them as a baseline for any extension of the Songer database. Secondly, in regards to machine learning, we show quite clearly that any input which does not include the complete opinion text in some form cannot result in a good overall performance. That is important as it shows that other aspects which are otherwise very useful in the domain of law, such as citations for citation networks, do not contain enough information for this specific task. This holds despite the fact that when using citations as input, the classifier uses many citations to which it assigns the correct ideology label if one were to label them by hand. However, when taken as an aggregation, neither citations nor quotations are distinctive enough. Moreover, while the Songer database features four labels, our results show that the error the classifier makes on the "mixed" label is nearly equally split between "conservative" and "liberal". As the "other" label is negligible in terms of occurrence, we can, therefore, conclude that training a classifier only on the two labels "conservative" and "liberal" does not introduce any systematic. Due to the increase in performance, such a setup should consequently be preferred. Lastly, looking at the regression results, it may be that text alone is not enough. Future research should therefore also think about taking meta-information, such as the circuit court it was heard at, into account. Moreover, looking at the literature of the median judge (e.g. Martin et al., 2004) it may also be important with which other judges a judge sits on a panel. This may be another important aspect, a machine learning classifier may have to take into account.

We hope that our work acts as a baseline on which future work can build on. The obvious next step is to scale back on the interpretability of the model in favor of sophistication: Specifically, we propose a modified doc2vec model in combination with an attention mechanism. Furthermore, future work could stack multiple classification algorithms tailored more closely to the rules of the coding book that the Songer database provides.

Another exciting avenue for future work is to compare in-depth the differences, advantages, and disadvantages of various methodological approaches. A particular exciting comparison is a Bayesian framework, as proposed by Martin and Quinn (2002), compared to machine learning approaches, as suggested by this paper.

Apart from methodological extensions, a more content-related one is particularly interesting: Most of the literature is targeted towards high ranking courts, such that the Supreme Court or Circuit Courts. This lack of attention towards lower courts might stem from the fact that the universe of cases to code is vast. Consequently, not even a partially coded data set, as far as political ideology labels are concerned, is available for lower courts. A classifier trained on Circuit Courts' opinions could predict the label for opinions of lower courts and, by that, help to close this particular gap in the literature.

CRediT authorship contribution statement

Carina I. Hausladen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Marcel H. Schubert:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Elliott Ash:** Conceptualization, Methodology, Resources, Data curation, Writing - review & editing, Supervision, Project administration.

Acknowledgements

We would like to thank Prof. Rok Spruk for his extensive feedback on statistics and evaluation methods, the members of the Max Planck Institute for Research of Collective Goods as well as the participants of the second Empirical Legal Studies Replication Conference for their comments, the ETH Center for Law & Economics for providing computing capabilities and additional data, and the Max Planck Computing and Data Facility for use of their cluster and for supporting us in all technical matters. This work was supported by an IPAK-Travel-Grant, by the Max-Planck Institute for Research on Collective Goods as well as the University of Cologne.

Appendix A. Replication

Table 5
Overview of all tables and figures in Landes and Posner (2009) dealing with the circuit courts.

Analysis of court of appeals voting: 1925–2002	
Table 11	Court of Appeals Votes by Subject Matter and Ideology for 538 Court of Appeals Judges Only: 1925–2002
Figure 3	Total Votes by Year Appointed to the Court of Appeals
Table 12	Fraction of Mixed (M), Conservative (C) and Liberal (L) Votes for 538 U.S. Court of Appeals Judges by President at Time of Appointment: 1925–2002
Table 13	Regression Analysis of Court of Appeals Votes: 1925–2002 (*t*-statistics in parentheses)
Table 14	Regression Analysis of Court of Appeals Votes: 1960–2002 (*t*-statistics in parentheses)
Table 15	Circuit Effects on Ideology of Judges' Votes
Table 16	Regression Analysis of Appellate Court Votes: Current Judges (*t*-statistics in parentheses)

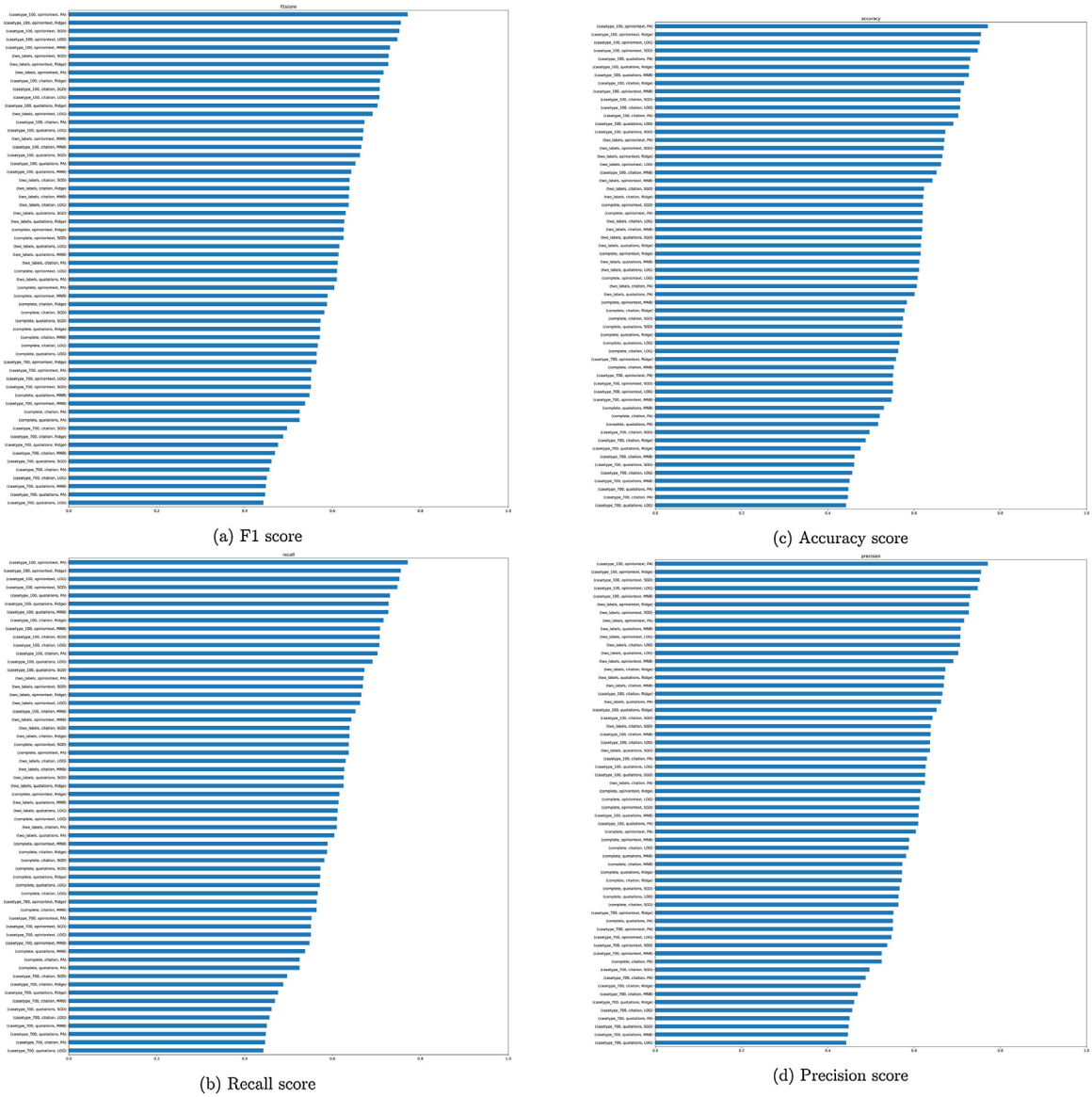
Appendix B. Data pre-processing

We applied pre-processing tailored to our data. As we use data from Lexis, each opinion had a specific structure. We extracted the text and split it into parts when encountering more than a single newline character. Special characters such as 'newline'-characters and roman numbers were removed.

If a potential heading was found within the text, we excluded it. The reason being that such a heading would potentially include biasing information such as judge names. It is especially important to exclude those, as the model could focus on judge names as a proxy for the directionality as most cases were decided without dissent. This is an issue in our empirical context because we would like to use the predicted data to analyze judge characteristics. Including the judges in the prediction would induce mechanical correlation.

In a second step, we applied regular expressions trying to capture the part of the opinion in which judges might dissent from the majority. Including a dissenting part which by its nature goes against the directionality of the majority in the input would not only add noise but may also lead the classifier to average over the different directions, leading to an overall worse performance. If we found a dissent, we split off the relevant paragraph and saved it as an extra entry in the database, marking it as 'dissent'. We excluded those entries and did not use them as input.

Appendix C. All classifier input combinations



(a) F1 score (c) Accuracy score

(b) Recall score (d) Precision score

Fig. 10. Various performance metrics for all different combinations tested.

Appendix D. Judges

Tables 6–9 present yet another way how to assess the performance of the best classifier. We predict the directionality of an opinion and use it to calculate the fraction of conservative or liberal votes by a judge. We split the population of judges by the party of the appointing president, resulting in four different specifications. Overall, actual and predicted fractions of votes by the ten highest ranked judge by the specification are pretty similar and reassures that our classifier performs sufficiently well for our analysis.

Table 6
10 judges with highest fraction of conservative votes, appointed by conservative presidents.

Frac con	Sum	Name	Frac con	Sum	Name
0.89	48	Barksdale, Rhesa H.	0.87	48	Barksdale, Rhesa H.
0.85	69	Loken, James B.	0.87	69	Loken, James B.
0.84	65	Hansen, David R.	0.83	66	Arnold, Morris S.
0.83	110	Easterbrook, Frank H.	0.82	109	Easterbrook, Frank H.
0.82	28	O'Scannlain, Diaruid F.	0.80	15	Lewis, Robert E.
0.82	61	Luttig, J. Michael	0.80	65	Hansen, David R.
0.80	93	Edmondson, James L.	0.80	44	DeMoss, Harold R., Jr.
0.80	72	Magill, Frank J.	0.79	61	Jones, Edith H.
0.80	104	Boudin, Michael	0.79	103	Boudin, Michael
0.80	45	DeMoss, Harold R., Jr.	0.78	97	Higginbotham, Patrick E.
Note:		Hand-labelled data	*Note:*		Predicted data

Table 7
10 judges with highest fraction of liberal votes, appointed by conservative presidents.

Frac lib	Sum	Name	Frac lib	Sum	Name
0.71	11	Thomas, Clarence	0.63	44	Hitz, William
0.63	44	Hitz, William	0.62	11	Thomas, Clarence
0.59	137	Gibbons, John J.	0.57	119	Wilbur, Curtis D.
0.58	39	Waddill, Edmund, Jr.	0.56	116	Van Orsdel, Josiah A.
0.58	46	Miller, William Ernest	0.56	70	Thompson, Joseph W.
0.58	73	Mansmann, Carol Los	0.56	46	Miller, William Ernest
0.58	43	Pratt, George C.	0.56	55	Roth, Jane R.
0.56	56	Roth, Jane R.	0.56	142	Northcutt, Elliott
0.56	142	Northcutt, Elliott	0.56	108	Lively, Frederick P.
0.56	107	Lively, Frederick P.	0.55	43	Pratt, George C.
Note:		Hand-labelled data	*Note:*		Predicted data

Table 8
10 judges with highest fraction of conservative votes, appointed by liberal presidents.

Frac con	Sum	Name	Frac con	Sum	Name
0.89	45	Evans, Terence Thomas	0.82	44	Evans, Terence Thomas
0.84	38	Parker, Robert Manley	0.81	37	Parker, Robert Manley
0.78	69	Williams, Jerre S.	0.80	20	Rutledge, Wiley Blount
0.76	83	Garza, Reynaldo	0.78	27	King, Carolyn Dineen
0.75	60	Anderson, Robert P.	0.76	82	Garza, Reynaldo
0.74	27	King, Carolyn Dineen	0.75	134	Breyer, Stephen G.
0.74	78	Mehaffy, Pat	0.74	163	McMillian, Theodore
0.73	131	Miller, Wilbur K., Jr.	0.74	19	Cole, Ransey Guy, Jr.
0.73	37	Murphy, Michael R.	0.74	68	Williams, Jerre S.
0.73	11	Kravitch, Phyllis A.	0.73	30	Stewart, Carl Edmond
Note:		Hand-labelled data	*Note:*		Predicted data

Table 9
10 judges with highest fraction of liberal votes, appointed by liberal presidents.

Frac lib	Sum	Name	Frac lib	Sum	Name
0.71	11	Faris, Charles	0.66	24	Russell, Robert L.
0.71	11	Thomas, Sidney Runyan	0.63	14	Sarokin, Haddon Lee
0.67	16	Hough, Charles M.	0.63	22	Strum, Louie
0.66	24	Russell, Robert L.	0.62	27	O'Connell, John J.
0.66	51	Haney, Bert E.	0.62	24	Clark, William
0.65	29	Ferguson, Warren J.	0.61	16	Hough, Charles M.
0.63	99	Higginbotham, Aloyisus Leon	0.61	98	Higginbotham, Aloyisus Leon
0.63	14	Sarokin, Haddon Lee	0.60	150	Robinson, Spottswood W., III
0.63	22	Strum, Louie	0.60	51	Haney, Bert E.
0.62	24	Clark, William	0.57	31	Lucero, Carlos
Note:		Hand-labelled data	*Note:*		Predicted data

Appendix E. Robustness checks

Additionally to the histograms found in Fig. 9, we go on to analyze the EBA's statistics on civil cases, displayed by Table 10a.

For civil cases, we estimated 510 regression models. Fig. 9a provides information about the share of regression coefficients that are statistically significant as well as lower (column 1) or greater (column 2) than zero. There was no coefficient significant for which the size of at least 50 percent of estimated coefficients lies below zero. By contrast, there were three coefficients found to be significant while having values larger than zero in at least 50 percent of the estimated models. These were the fraction of republican senators at the point of election (92 percent), the fraction of miscellaneous votes (64 percent) as well as circuit 1 (100 percent). Consequently, Leamer (1985)'s EBA (column 3), defines circuit 1 as the only robust variable. Furthermore, Table 10a includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). Fig. 9a suggests that a normal distribution does not sufficiently well approximate the regression coefficients' distribution. For this reason, we focus on Sala-i-Martin (1997) EBA results from a model that does make assumptions about the coefficients' distributions. As a rule of thumb, those variables for which more than 90 percent of the regression coefficients' cumulative distribution is located either above or below zero, can be interpreted as being robustly connected with the dependent variable (Hlavac, 2016). For the variables of being black (96 percent), the years of having served as a district court judge (93 percent), as well as for the fraction of economic votes (93 percent), more than 90 percent of the cumulative distributions lie below zero. By contrast, for the variables of being appointed by a conservative president (99 percent), the fraction of miscellaneous votes (98 percent) as well as for circuit 1 (100 percent), more than 90 percent of the cumulative distributions lie above zero.

EBA statistics for criminal cases, displayed in Table 10b, are interpreted below. Overall, 127 regression models were estimated. Columns 1 and 2 of Table 10b show the fraction of the respective regression coefficients that are statistically significant and lower or greater than zero at the same time. Only for the dummy variable *Black*, more than 88 percent of the values estimated were significant and smaller than zero. By contrast, there were three coefficients, *Pres* (100 percent), *circuit 8* (100 percent) and *circuit 10* (100 percent) found to be significant and showing more than 50 percent of its values larger than zero. Table 10b summarizes results from Leamer (1985)'s EBA (column 3). This test concludes that three variables are found to be robustly connected with the dependent variable, which are *Pres* as well as *circuits 8* and *10*. Fur-

Table 10
Extreme bounds analysis.

	β sign & <0	β sign & >0	leamer robust	cdf β <=0 generic	cdf β > 0 generic
Civil cases					
(Intercept)	0.25	0.50	FALSE	0.47	0.53
Pres	0.00	0.92	FALSE	0.01	0.99
SenRep	0.00	0.00	FALSE	0.30	0.70
YrAppt	0.00	0.50	FALSE	0.11	0.89
Gender	0.00	0.00	FALSE	0.33	0.67
Black	0.47	0.00	FALSE	0.96	0.04
DistrictCourt	0.01	0.00	FALSE	0.93	0.07
FracEcon	0.50	0.00	FALSE	0.95	0.05
FracMisc	0.00	0.64	FALSE	0.02	0.98
CircuitVariables1	0.00	1.00	TRUE	0.00	1.00
CircuitVariables2	0.00	0.00	FALSE	0.52	0.48
CircuitVariables3	0.00	0.00	FALSE	0.91	0.09
CircuitVariables4	0.00	0.00	FALSE	0.62	0.38
CircuitVariables5	0.00	0.00	FALSE	0.29	0.71
CircuitVariables6	0.00	0.00	FALSE	0.42	0.58
CircuitVariables7	0.00	0.00	FALSE	0.08	0.92
CircuitVariables8	0.00	0.00	FALSE	0.21	0.79
CircuitVariables9	0.00	0.00	FALSE	0.83	0.17
CircuitVariables10	0.00	0.00	FALSE	0.71	0.29
CircuitVariables11	0.00	0.00	FALSE	0.43	0.57
criminal cases					
(Intercept)	0.00	0.50	FALSE	0.14	0.86
Pres	0.00	1.00	TRUE	0.00	1.00
SenRep	0.00	0.00	FALSE	0.70	0.30
YrAppt	0.00	0.00	FALSE	0.44	0.56
Gender	0.00	0.00	FALSE	0.61	0.39
Black	0.88	0.00	FALSE	0.99	0.01
DistrictCourt	0.00	0.00	FALSE	0.78	0.22
CircuitVariables1	0.00	0.00	FALSE	0.06	0.94
CircuitVariables2	0.00	0.00	FALSE	0.60	0.40
CircuitVariables3	0.00	0.00	FALSE	0.71	0.29
CircuitVariables4	0.00	0.00	FALSE	0.46	0.54
CircuitVariables5	0.00	0.00	FALSE	0.25	0.75
CircuitVariables6	0.00	0.00	FALSE	0.49	0.51
CircuitVariables7	0.00	0.00	FALSE	0.07	0.93
CircuitVariables8	0.00	1.00	TRUE	0.01	0.99
CircuitVariables9	0.00	0.00	FALSE	0.33	0.67
CircuitVariables10	0.00	1.00	TRUE	0.01	0.99
CircuitVariables11	0.00	0.00	FALSE	0.14	0.86

thermore, Table 10b includes results from Sala-i-Martin (1997)'s EBA (columns 4 and 5). As was the case with civil cases, Fig. 9b suggests that a normal distribution does not fit the coefficients' distribution very well. For this reason, we focus on EBA results from a parameter-free model. For *Black* (99 percent), more than 90 percent of the cumulative distributions lie below zero. By contrast, for the variables of being appointed by a conservative president (*Pres*) (100 percent), for circuit 1 (94 percent), circuit 7 (93 percent), circuit 8 (99 percent) and circuit 10 (99 percent) more than 90 percent of the cumulative distributions lie above zero.

References

Abadie, A., 2018. Statistical Non-Significance in Empirical Economics., pp. 1–25, December.

Aletras, N., et al., 2016. Predicting judicial decisions of the European court of human rights: a natural language processing perspective. PeerJ Comput. Sci. 2016 (10), e93.

Ash, E., Chen, D.L., 2018. Mapping the geometry of law using document embeddings. SSRN Electron. J.

Ash, E., Chen, D.L., Lu, W., 2018. Motivated Reasoning in the Field: Partisanship in Precedent, Prose, Vote, and Retirement in US Circuit Courts, 1800–2013, Technical Report.

Boella, G., Di Caro, L., Humphreys, L., 2011. Using classification to support legal knowledge engineers in the Eunomos legal document management system. Fifth International Workshop on Juris-Informatics (JURISIN).

Cameron, C., Gelbach, J., Miller, D., 2006. Robust Inference with Multi-Way Clustering., pp. 1–34, NBER Working Paper, September.

Cao, Y., Ash, E., Chen, D.L., 2018. Automated fact-value distinction in court opinions. SSRN Electron. J.

Casper, G., Posner, R.A., 1974. A study of the supreme court's caseload. J. Legal Stud. 3 (2), 339–375.

Chandler, S.J., 2005. The Network Structure of Supreme Court Jurisprudence. University of Houston Law Center, (2005-W):1.

Crammer, K., et al., 2006. Online passive-aggressive algorithms. J. Mach. Learn. Res. 7, 551–585.

Dainow, J., 1966. The civil law and the common law: some points of comparison. Am. J. Comp. Law 15, 419.

Epstein, L., Martin, A.D., et al., 2012. Ideology and the study of judicial behavior. In: Ideology, Psychology & Law, 705.

Epstein, L., Segal, J.A., 2005. Advice and Consent: The Politics of Judicial Appointments.

Fong, C., Tyler, M., 2018. Machine Learning Predictions as Regression Covariates.

Giles, M.W., Hettinger, V.A., Peppers, T., 2001. Picking federal judges: a note on policy and partisan selection agendas. Polit. Res. Q. 54 (3), 623–641.

Ginn, M.H., Searles, K., Jones, A., 2015. Vouching for the court? How high stakes affect knowledge and support of the supreme court. Justice Syst. J. 36 (2), 163–179.

Hlavac, M., 2016. ExtremeBounds: extreme bounds analysis in R. J. Stat. Softw. 72 (9).

Johnson, S.W., Songer, D.R., Jilani, N.A., 2011. Judge gender, critical mass, and decision making in the appellate courts of Canada. J. Women Polit. Policy 32 (3), 237–260.

Kassow, B., Songer, D.R., Fix, M.P., 2012. The influence of precedent on state supreme courts. Polit. Res. Q. 65 (2), 372–384.

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification, Technical report.

Landes, W.M., Posner, R.A., 2009. Rational judicial behavior: a statistical study. J. Legal Anal. 1 (2), 775–831.

Lauderdale, B.E., Clark, T.S., 2014. Scaling politically meaningful dimensions using texts and votes. Am. J. Polit. Sci. 58 (3), 754–771.

Lauderdale, B.E., Herzog, A., 2016. Measuring political positions from legislative speech. Polit. Anal., 1–21.

Laver, M., Benoit, K., Garry, J., 2003. Extracting policy positions from political texts using words as data. Am. Polit. Sci. Rev. 97 (2), 311–331.
Leamer, E.E., 1985. Sensitivity analyses would help. Am. Econ. Rev. 75 (3), 308–313.
Martin, A.D., Quinn, K.M., 2001. The Dimensions of Supreme Court Decision Making: Again Revisiting The Judicial Mind., pp. 1–37.
Martin, A.D., Quinn, K.M., 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. Polit. Anal. 10 (2), 134–153.
Martin, A.D., Quinn, K.M., Epstein, L., 2004. The Median Justice on the United States Supreme Court, Technical Report.
Masood, A.S., Songer, D.R., 2013. Reevaluating the implications of decision-making models. J. Law Courts 1 (2), 363–389.
McFadden, D., 1973. Conditional Logit Analysis of Qualitative Choice Behavior.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (October), 2825–2830.
Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. 10 (3), 61–74.
Randazzo, K.A., Waterman, R.W., Fix, M.P., 2010. State supreme courts and the effects of statutory constraint. Polit. Res. Q. 64 (4), 779–789.
Reid, R., Randazzo, K.A., 2016. Statutory language and the separation of powers. Just. Syst. J. 37 (3), 246–258.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you?: explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
Rifkin, R., Lippert, R., 2007. Notes on Regularized Least Squares. Massachusetts Institute of Technology, Cambridge, Technical Report.
Sala-i-Martin, X.X., 1997. I Just Ran Four Million Regressions.
Schmidt, M., Le Roux, N., Bach, F., 2017. Minimizing finite sums with the stochastic average gradient. Math. Program. 162 (1–2), 83–112.
Segal, J.A., Cover, A.D., 1989. Ideological values and the votes of US supreme court justices. Am. Polit. Sci. Rev. 83 (2), 557–565.
Shrestha, P., et al., 2017. Convolutional neural networks for authorship attribution of short texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: vol. 2, 669–674, Short Papers.
Sidorov, G., et al., 2014. Syntactic n-grams as machine learning features for natural language processing. Expert Syst. Appl. 41 (3), 853–860.
Slapin, J.B., Proksch, S.-O., 2008. A scaling model for estimating time-series part positions from texts. Am. J. Polit. Sci. 52 (3), 705–722.
Songer, D.R., 1993. The United States Court of Appeals Database – Documentation for Phase I.
Sturm, H.P., Pritchett, H.C., 2006. The roosevelt court, a study in judicial politics and values, 1937–1947. West. Polit. Q. 2 (3), 465.
Suen, C.Y., 1979. N-gram statistics for natural language understanding and text processing. IEEE Trans. Pattern Anal. Mach. Intell. 2, 164–172.
Sulea Octavia, M., et al., 2017. Exploring the use of text classification in the legal domain. CEUR Workshop Proceedings, 2143.
Undavia, S., Meyers, A., Ortega, J., 2018. A comparative study of classifying legal documents with neural networks. In: Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, October, pp. 515–522.
Vaswani, A., et al., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems., pp. 5998–6008.
Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 919–926.