Economica

Economica (2022) **89**, 362–376
doi:10.1111/ecca.12397

Measuring Judicial Sentiment: Methods and Application to US Circuit Courts

By ELLIOTT ASH*, DANIEL L. CHEN† and SERGIO GALLETTA‡

**ETH Zürich* †*Toulouse Institute for Advanced Study* ‡*University of Bergamo*

Final version received 14 September 2021.

This paper provides a general method for analysing the sentiments expressed in the language of judicial rulings. We apply natural language processing tools to the text of US appellate court opinions to extrapolate judges' sentiments (positive/good vs. negative/bad) towards a number of target social groups. We explore descriptively how these sentiments vary over time and across types of judges. In addition, we provide a method for using random assignment of judges in an instrumental variables framework to estimate causal effects of judges' sentiments. In an empirical application, we show that more positive sentiment influences future judges by increasing the likelihood of reversal but also increasing the number of forward citations.

INTRODUCTION

Law is composed of natural language, therefore understanding its effects quantitatively has remained elusive for researchers using the standard empirical toolkit (Ash and Chen 2019). An important dimension of legal language is its *sentiment*—that is, its positive or negative tone. Does a more optimistic tone make a judge more persuasive? Or instead is a more critical tone more effective? This paper provides methods for estimating judicial sentiment and analysing its impacts on other judges and the path of the law.

A first contribution of this paper is the method used to infer judges' preferences towards specific target groups (e.g. black, white, Republicans and Democrats). Rather than focusing on the direction of decisions (for/against a particular group), we apply natural language processing techniques to the text of US Circuit Court opinions. In particular, we draw on recent embedding methods, which vectorize words and documents in a relatively low-dimensional space, where locations and directions encode meanings and associations. At a sentence level, our algorithm measures both the relevance to each of the different groups and the level of sentiment (positive/warm or negative/cold). From these sentence-level measures, we compute the relative sentiment in a case by the correlation between group associations and sentiment associations. This flexible and informative solution to measuring judicial attitudes highlights the growing literature using text to understand biases and preferences (Caliskan *et al.* 2017). Our paper is the first to apply these methods to judicial opinions to analyse their legal impact.

The paper's second contribution is to address the empirical challenge that judge sentiments do not vary randomly over time and space and therefore this variable is likely to be endogenously determined in many contexts. Unlike the literature that instruments for judicial decisions using judge leniency (e.g. Galasso and Schankerman 2014; Dobbie and Song 2015; Sampat and Williams 2019), there is no straightforward way to instrument for sentiment expressed in text. We apply machine learning tools to extract predictive power in the first stage from a high-dimensional set of instruments describing the biographical characteristics of judges assigned to these cases. Our approach extends the literature on sparse optimal instruments using cross-fitting techniques (Belloni *et al.* 2012; Chernozhukov *et al.* 2017; Chen *et al.* 2020). Specifically, we apply elastic net

regression to the standardized judge characteristics and construct cross-validated instruments using out-of-fold data. The predictions from these estimates are then gathered together to be used as instruments in the second stage.

To illustrate the usefulness of our method, we do two things. First, we provide descriptive evidence about the variation in expressed sentiment across time, across circuits, and across different types of judges (e.g. whether appointed by Democrat/Republican President, age, gender and race). We show that sentiment is relatively stable across time and space while varying across groups of judges. For example, we find that sentiment toward African-Americans is lower for white, male, Republican judges. We show that judge writing sentiment is negatively correlated with the expressed sentiments in surveys toward the same social groups. We also demonstrate some limits on language sentiment measures.

Second, we apply the instrumental variables approach to test whether the sentiments expressed in judicial rulings have actual consequences in the development of the law. In particular, we show that more positive/warm (rather than negative/cold) case sentiment increases the likelihood that the Supreme Court reviews the Circuit Court decision, and the chances that the decision is eventually reversed. Moreover, we find that expressed sentiment increases the probability of an opinion being cited in subsequent cases. Expressed sentiment in judicial opinions matters for the responses of other judges and for the path of the law.

This paper adds to the emerging literature using machine learning methods to overcome limitations of standard datasets—in our case, isolating variation in judicial sentiments. Several papers in political economics have used supervised learning to extract measures of partisanship from text (Gentzkow and Shapiro 2010; Ash *et al*. 2017; Gentzkow *et al*. 2019). Meanwhile, unsupervised learning algorithms have been used to extract measures of individual behaviours (Bandiera *et al*. 2020) and attitudes (Draca and Schwarz 2019) from high-dimensional data. Ash and Chen (2017) use embeddings to perform a descriptive analysis of legal language. Kozlowski *et al*. (2019) use word embedding models to study the historical evolution of the culture understandings of social classes, by analysing millions of books published over 100 years.

Methodologically, our approach to estimate causal effects is close to that of Belloni *et al*. (2012) and Chernozhukov *et al*. (2017) as we use machine learning techniques to account for sparsity in the potential set of instruments. Moreover, we build on existing studies that exploit random assignment of judges for identification (Di Tella and Schargrodsky 2013; Galasso and Schankerman 2014; Kling 2006; Maestas *et al*. 2013).

The remainder of the paper is organized as follows. Section I describes the institutional background and data. Section II describes the method use to measure judges' sentiment. Section III provides descriptive evidence about the variation in expressed sentiment. Section IV details the instrumental variables approach. Section V reports an application of our methodology. Section VI concludes.

I. Institutional Setting and Data

The US Federal Courts system

The US Federal Courts system is organized on three levels: the national level (Supreme Court), intermediate level (Circuit Courts) and local level (District Courts). Our analysis exploits features that are specific to the intermediate level. There are 11 regional US Circuit Courts. Each of these courts is responsible for 3–9 states (see Figure A.1 of the

Online Appendix), in addition to a court for the District of Columbia (DC) and a Federal Circuit Court that has national jurisdiction in specific domains. All Circuit Court judges are appointed for life. On average, a Circuit has 17 judges, with a minimum of 8 and maximum of 40. Each case is assigned to a panel of three of the Circuit's judges.

The Circuit Courts play a crucial policy-making role in US law because their judges review the decisions taken by the District Courts. A large majority of appeals terminate at this stage, and those decisions are binding precedent within the circuit. Therefore judicial decisions have the force of law, and become official articulations of legal and social norms.

Unsurprisingly, then, these decisions and the associated opinions are the target of significant attention by the media and the public (Bromley 1994). Evidence of substantial public attention to court opinions includes Weinrib (2012), who documents the response by ACLU attorneys to major Circuit Court decisions on free speech. The attorneys responded by mobilizing people in the media in favour of stronger free-speech protections. Clark *et al*. (2018) find significant responses on Twitter after several court decisions. Lim *et al*. (2015) document the frequent coverage of criminal decisions in newspapers.

Data

We have assembled data from a range of sources. To create the judges' sentiment measure, we use the complete collection of US Courts of Appeals opinions from 1961 to 2013. The corpus includes all published cases and comes from Bloomberg Law. For the empirical application, we use some additional case-level metadata. This includes the direction of the decision (affirm/reverse), whether the Supreme Court reviewed the case, if the Supreme Court reversed the decision, the number of citations, and general category labels.[1]

Further, we collected the biographical information of judges working in the Circuit Courts during this period. We match each judge with data from the Federal Appeals and District Court Attribute Data.[2] We integrate this information with data from the Federal Judicial Center's biographies of judges and previous data collection (Chen *et al*. 2016). Overall, we have a total of 60 variables that refer to judges' biographical characteristics that we use to support the proposed empirical methodology. These variables include, for instance: age, geographic history, education, occupational history, governmental positions, military service, religion, race, gender and political affiliation.

II. A Measure of Judicial Sentiment

To measure judicial sentiment, we apply an embedding model to the text of US Circuit Court opinions. Embedding models are a recent neuro-linguistic programming technique that has been mainly implemented in computational linguistics for prediction tasks (e.g. Mikolov *et al*. 2013). For example, embedding methods are used to predict the next word in an incomplete sentence. During the training process, the algorithm assigns each word to a vector in a shared geometric space. This procedure allows words to cluster near semantically similar words. In consequence, the position in the space encodes the context in which words are used. The closer two words are located in the language space, the higher the similarity of the context. Moreover, trained embeddings encode meaningful information about analogies.[3]

In a nutshell, our approach exploits vector similarity measures (i.e. cosine similarity) to evaluate the sentiment expressed by judges in each case (positive vs. negative) as well as the degree to which a case is about specific pre-selected target groups (e.g. women, business, Republicans). This idea is closely related to the work of Caliskan *et al*. (2017) and Kozlowski *et al*. (2019), who use word space to gauge biased associations in text. By using tools similar to ours applied to millions of books, they study the evolution of culture over the last 100 years.

We use these tools to explore another important structure in culture: the law. Caliskan *et al*. (2017) and Kozlowski *et al*. (2019) focus on gender and class, while we focus on a broader range of groups that are salient in legal disputes. Our contribution is more empirically oriented because we look at the impacts of language variation on the law and society.

Concretely, the starting point for measuring judicial sentiment is the collection of Circuit Court opinions. First, we parse the raw text into Python and use the Python module *nltk* to tokenize sentences. Next, we map sentences into vectors using the Python module Doc2Vec (Mikolov *et al*. 2013; Le and Mikolov 2014). This algorithm represents words and sentences in a shared vector space (in our case, 200 dimensions). As already mentioned, words that tend to have similar contexts are located near each other (we used a window size of 5).[4] Similarly, sentences with comparable language tend to locate close to each other and tend to locate close to words contained in the sentence. Dai *et al*. (2015) illustrate the use of Doc2Vec to analyse similarities and analogical relations between documents (see also Ash and Chen 2019).

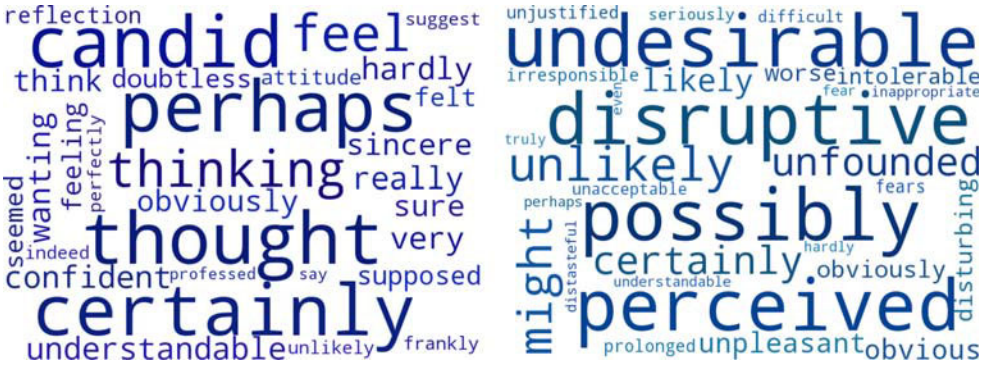
As we want to measure judges' sentiment towards specific groups/ideas, we would like a set of target groups that is standard in opinion surveys over a long time period. We use the categories assessed in the feeling thermometer questions of the American National Election Survey (ANES).[5] With the trained Doc2Vec model in hand, we obtain vectors for 19 of the ANES targets as the average of a set of words for each target (see Section A.1 of the Online Appendix). Blacks, for example, are identified from black, blacks, african, africans, african-american, african-americans, negro and negroes. Figure A.2 of the Online Appendix shows word clouds for the words most associated with each target.

In the case law corpus, we compute the cosine similarity of each sentence vector to each of the targets.[6] The cosine similarity metric provides an estimate of semantic association between each sentence with each specific target group. Formally, let W_{id}^{k} represent the similarity of sentence d in case i to target k. If needed, we represent the average similarity of a case i to target k as

$$W_{i}^{k} = \frac{1}{|D_i|} \sum_{d \in D_i} W_{id}^{k},$$

where D_i is the set of sentences in i.

Next, for each sentence we compute a metric for positive and negative sentiment. To construct the sentiment dimension, we use a dictionary of positive words (e.g. 'warm', 'favourable', 'good') and negative words (e.g. 'cold', 'unfavourable', 'bad') (see Section A.1 of the Online Appendix). Figure 1 shows the words most associated with the positive and negative attributes. Similarly to what was just described about the target groups, we find the average vector for these word sets, and then compute the cosine similarity of each sentence to the averaged sentiment vector. We define the sentiment S_{id}



(a) Positive sentiment (b) Negative sentiment

FIGURE 1. Positive and negative sentiment language. *Notes*: Most similar words in the embedding space to the average vector for the lexicon of positive words (left) and negative words (right). See text for details.

for sentence d in case i as the cosine similarity to the positive vector, minus the cosine similarity to the negative vector.

Finally, we aggregate these sentence-level statistics to the case level. We construct the case-level sentiment towards target k as

$$S_{ik} = \sum_{d \in D_i} S_{id} \cdot W_{id}^{k},$$

the dot product of these two vectors. These measures can be aggregated further by computing the average sentiment across cases. For example, let C_{ct} be the set of cases filed in circuit c during year t. Then we can define

$$\mathbf{S}_{ckt} = \frac{1}{|C_{ct}|} \sum_{i \in C_{ct}} \mathbf{S}_{ictk},$$

the average case-level sentiment towards k for cases in circuit–year ct.

III. DESCRIPTIVE EVIDENCE ON JUDGE SENTIMENT

This section investigates the details of our measure of judicial sentiment, looking at how it varies across different dimensions. First, we consider variation of expressed sentiment over time. Figure 2 shows the trend of sentiment towards each target group from 1961 to 2013. For most of the targets, the measure is stable. However, recognizable positive trends are present in judicial sentiment towards Business, Catholic and Democrat. Meanwhile, there is a negative trend for Liberal and Supreme Court. The increase in positive sentiment toward Business and negative sentiment towards Liberal could be part of a previously noted increasing economic conservatism in the judiciary (Ash *et al*. 2020). The positive trend for Democrat, while inconsistent with the conservatism interpretation, could be an artefact of our model being case-insensitive—there could be an increase in positive sentiment toward small-d democracy and democratic principles.

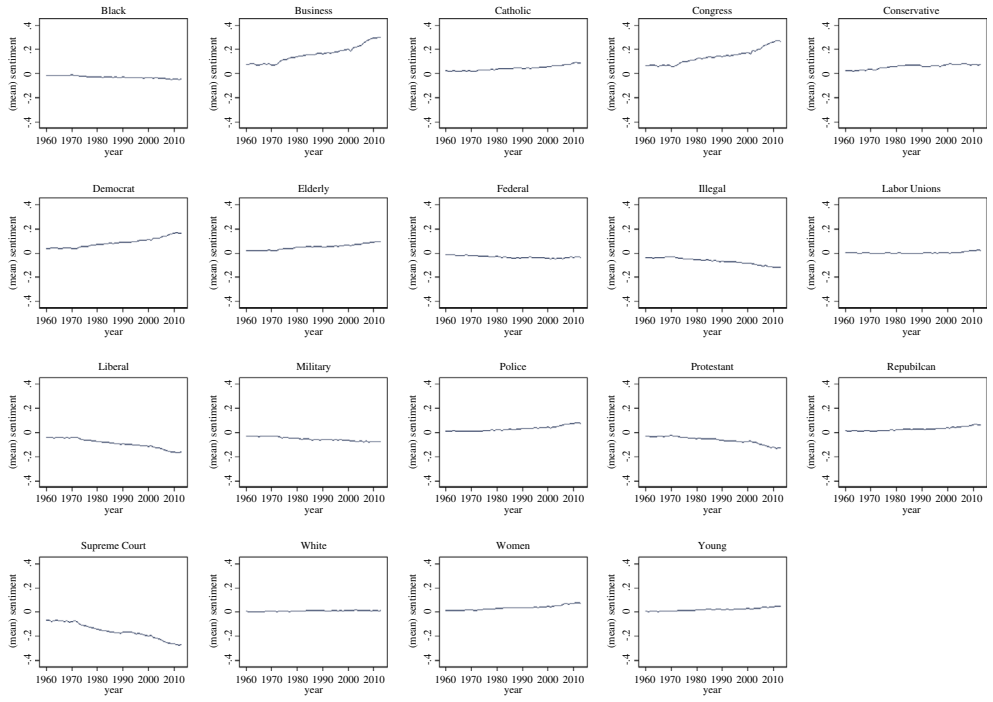


FIGURE 2. Judicial sentiment over time, by target.

Meanwhile, Figure A.4 of the Online Appendix shows variation across different Circuit Courts, each corresponding to a different geographical area. Here, we do not see large differences in our judicial measure across circuits for most targets. The DC Circuit (indicated as the 12th district in the figure) diverges in sentiment when the target groups are Labor Unions and Federal. These differences could be explained by the fact that this court covers cases that involve Congress and other government agencies, and therefore addresses issues that are different compared to the other courts.

Next, we consider how the characteristics of the judges assigned to a case relate to expressed sentiment. In Figure 3, we show the mean level of sentiment by target group residualizing the variable on court–year fixed effects beforehand, depending on the composition of the judge panel. In panel (a), we show the sentiment towards each group depending on whether the judges were appointed by a Democratic or a Republican president. In panel (b), we consider the gender composition of the group of judges, comparing all-male panels to ones with at least one female. Panel (c) looks at racial differences, comparing all-white panels to those with at least one non-white judge. In panel (d), we focus on age, comparing average sentiment expressed in rulings from a panel of judges whose members are all older than 50 years with those that have at least one judge who is younger than 50 years.

Overall, there are differences in sentiment depending on the demographic composition of the responsible judges. Some interesting patterns, for instance, are that for the Black/African-American target, the sentiment is generally negative, and relatively more negative when the judges are all Republicans, all males, all white, or older than 50 years. When the target is Business, there is generally a positive sentiment, and this is

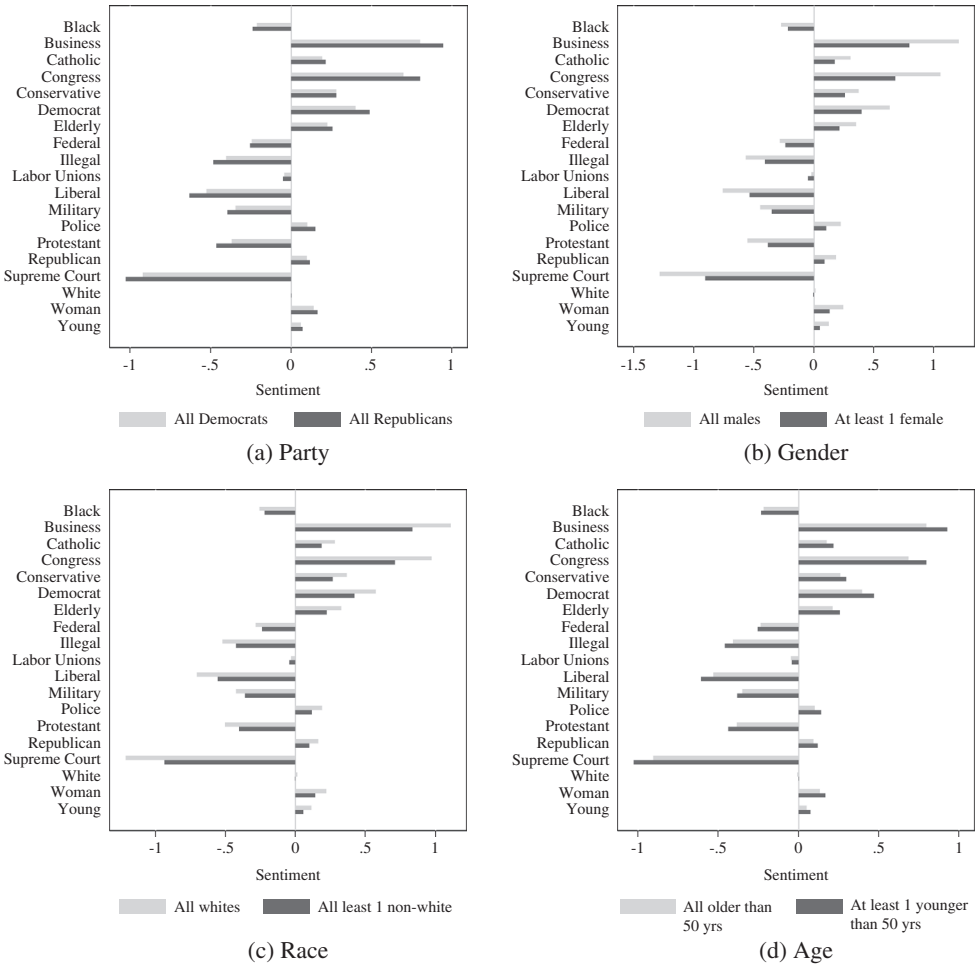


FIGURE 3. Demographic characteristics of the pool of judges. *Notes*: Judicial sentiment variation depending on the demographic characteristics of the assigned panel of judges.

higher when all judges are Republicans, all males, all white, and at least one judges is younger than 50 years.

Table A.2 of the Online Appendix shows that most of these differences are statistically significant. Yet these differences are not overwhelming in the sense that none of the comparisons indicate sentiments that have an opposite sign depending on the judge characteristic. And other comparisons are not that intuitive. For example, we find that when all judges are Republican, the sentiment is more positive towards Democrats. In addition, an all-male judge panel has more positive sentiment towards women. Overall, these results point out the potential limits that using text analysis can have in catching nuances in the rulings. Researchers should be cautious in the interpretation of this text-based evidence.

As a final descriptive exercise, we ask whether the sentiments expressed by judges in opinions are correlated with local circuit residents' sentiments reported in surveys. For this analysis we measure individuals' preferences towards the set of 19 target groups with information from the ANES. The feeling thermometer questions ask about attitudes

towards a specific target group by choosing a value from 0 to 100 (see Figure A.3 of the Online Appendix). A value closer to 100 reveals that the respondent feels warmly or favourably towards the target group, while closer to 0 means cold or unfavourable feelings towards the target group.

Figure 4 shows two binscatter diagrams for the relationship between judge and resident sentiment. In panel (a), we include year fixed effects interacted with target group fixed effects, showing that at any given time, the judge writing sentiments for a given target group are not correlated with resident reported sentiments in the cross-section. In panel (b), we include circuit fixed effects (interacted with target group fixed effects), showing that within-circuit changes in judge and resident sentiments are negatively correlated over time.

This negative correlation between judge sentiment and resident sentiment is not causal. There could be joint causality, or there could be a third confounding variable driving both measurements. To the extent that judges are influencing residents, rather than vice versa, our evidence is consistent with a backlash effect, where judicial rulings in favour of various groups trigger negative feelings by residents. Such a backlash effect would be consistent with previous work on abortion attitudes by Chen *et al*. (2016) and the evidence on gender equality in Wheaton (2020).

To summarize, our descriptive evidence presents a somewhat mixed picture of the relevance of judicial sentiment. While it does vary over time and is negatively correlated with changes in resident survey attitudes, it does not vary much by circuit or by judge identity. This mixed picture does not imply that sentiment does not matter, however. In the following sections, we turn to a causal analysis to see whether sentiment matters for how cases influence other judges.

IV. Estimating the Effect of Judicial Sentiment

In this section, we describe a framework to study the causal effects of judge sentiment in the context of the US Courts of Appeals. Typically, OLS regressions, even with fixed effects, would not produce causal evidence about sentiment. There could be confounders and joint causality, so the resulting estimates would likely be biased. To account for these

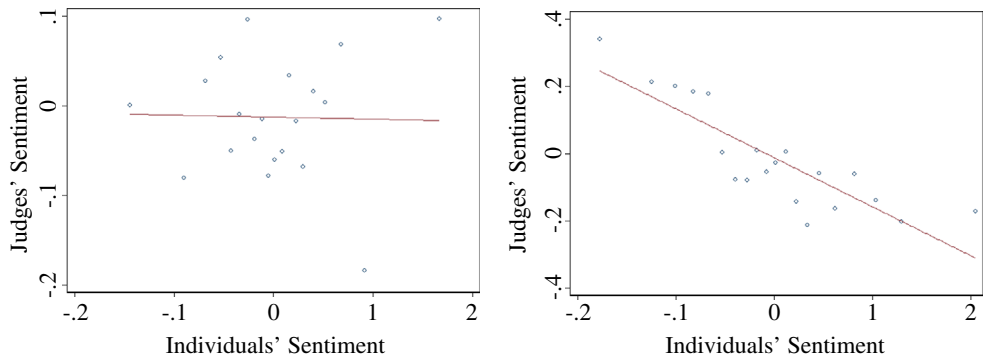


FIGURE 4. Correlation between judges' and individuals' sentiments. *Notes*: Binscatter diagrams displaying the correlation between judges' and individuals' sentiments. Panel (a) includes as controls year fixed effects × target fixed effects. Panel (b) includes as controls circuit fixed effects × target fixed effects.

endogeneity concerns, we propose an instrumental variables strategy that exploits the random assignment of judges to cases as a source of exogenous variation. In particular, we take advantage of the fact that judge characteristics are good cross-validated predictors of expressed sentiments, together with the fact that the personal characteristics of assigned judges to a case are as good as random once conditioned on their distribution in a given circuit–year.

Our approach combines identification features that are commonly adopted under random judge assignment with emerging machine learning methods. Specifically, we suggest the use of regularized regression to construct instruments from cross-validated predictions that are based on judges' characteristics. The methodology that we propose is not too distant from the ones already applied in the literature that use either a jackknife IV (see, for example, Dobbie and Song 2015; Kling 2006; Galasso and Schankerman 2014) or a split-sample two-stage IV (see, for example, Sampat and Williams 2019) to exploit judges' leniency variation. Our cross-validated prediction approach is similar to the split-sample two-stage IV methods proposed by Angrist and Krueger (1995) given that also in our case the instrument is constructed based on coefficients trained on out-of-fold data (see also Chen *et al*. 2020).

As a first step, we assign judge characteristics to cases and then to topics. Let $\mathbf{X}_{ict}$ be the average characteristics for the three judges assigned to case i in circuit c during year t. Then

$$\mathbf{J}_{ickt} = \mathbf{X}_{ict} \times W_i^k \tag{1}$$

is the vector of judge characteristics, weighted by the similarity to target k of the cases to which the judges are assigned.

As already noted, $\mathbf{J}_{ickt}$ contains a large number of characteristics (60 of them). We draw on recent developments in machine learning, to extract more predictive power from the estimates while avoiding over-fitting (see, for example, Chernozhukov *et al*. 2017; Chen *et al*. 2020). Specifically, to predict sentiment using the judges' characteristics, we can use regularized regression models such as LASSO, ridge regression, or elastic net. Next, we form the cross-validated prediction. The predicted endogenous regressor is the instrument in our two-stage least-squares (2SLS) regressions (Z_{ickt}).

We can now define the first-stage equation as

$$S_{ickt} = \gamma_k + \gamma_{ct} + \gamma_Z \, Z_{ickt} + \eta_{ickt}, \tag{2}$$

where S_{ickt} is the sentiment towards target k in a case i published in circuit c during year t. Z_{ickt} is the machine-learning-predicted instrument. γ_{ct} is a set of dummy variables (fixed effects) for each circuit–year, and γ_k is a set of dummy variables (fixed effects) for each target. η_{ckt} is the error term.

The second-stage estimating equation is

$$Y_{ickt} = \alpha_k + \alpha_{ct} + \beta \, \hat{S}_{ickt} + \varepsilon_{ickt}, \tag{3}$$

where the α terms are fixed effects, as previously defined. $\hat{S}_{ickt}$ is the predicted target sentiment as computed from the first stage—equation (2). Y_{ickt} is the outcome variable, and β is our coefficient of interest, giving the average effect of judge writing sentiment.

V. Empirical Application

In this section, we provide an application of our methodology by studying the relationship between judges' sentiment expressed in their rulings and the impact of their decisions on the law. This is relevant in a common-law system where judicial rulings provide precedent for future judges. Yet judges have a choice among many precedents, so they may select higher- or lower-sentiment cases. Thus the evidence speaks to the linguistic factors that judges find persuasive.

Specifically, we study the effect of expressed sentiment on the appeal outcomes of a case and how often it is cited. These provide measures of legal impact (e.g. Ash and MacLeod 2021). These measures include the likelihood that the Supreme Court reviews a Circuit Court decision, the likelihood that the Supreme Court reverses a Circuit Court decision, and the total number of citations to the Circuit Court decision. For this analysis, our observations are aggregated at the case–target level.

We begin our application following the procedure from Section IV to construct the instrumental variable from judges' characteristics. To prepare the data for the prediction task from which the instrument originates, we standardize to variance 1 the average judges' characteristics $\mathbf{J}_{ickt}$, as well as the judicial sentiment S_{ickt}, by target group. To create the instrument, we use elastic net. Elastic net is a linear regression with a penalized cost function to shrink coefficients toward zero and avoid over-fitting (Zou and Hastie 2005). The predictions are then formed using a fivefold cross-validation. We learned the cost-minimizing penalties: L1 = 0.2, L2 = 0.8, and a general penalty $\lambda = 0.00013$.[7] This means that in our data, the elastic net gives more weight to the ridge regression component than the LASSO component, while selecting a mild penalty.[8]

Next, we implement the first-stage equation (2) and we confirm that the instrument is strongly predictive of sentiment: coeff = 0.931, S.E. = 0.056, with F-statistic 278. But the instrument and endogenous regressor are far from being collinear ($R^2 = 0.006$), as shown in the scatterplot of Figure 5.

Table 1 presents the main findings for how judicial sentiment affects case appeal and citations. We report both OLS estimates and 2SLS estimates using the proposed constructed instrument, always including the relevant set of fixed effects and controls.

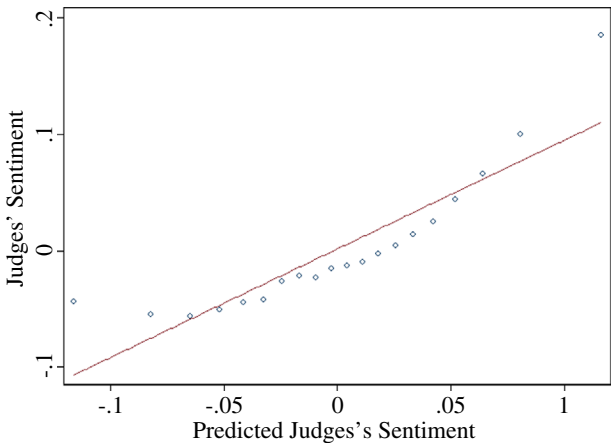


FIGURE 5. First-stage relationship. *Notes*: Binscatter diagram for the first-stage relationship (coeff = 0.931, S.E. = 0.056, $R^2 = 0.006$).

TABLE 1
EFFECT OF SENTIMENT ON CASE APPEAL AND CITATIONS

	Supreme Court reviewed		Supreme Court reviewed		Number of citations	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Judges' sentiment	0.002***	0.026***	0.001***	0.011***	0.111***	2.227***
	(0.000)	(0.003)	(0.000)	(0.002)	(0.016)	(0.535)
F-statistic		278		278		278
Observations	3,377,250	3,377,250	3,377,250	3,377,250	3,377,250	3,377,250

Notes
The dependent variable is a dummy variable identifying if a case has been reviewed by the Supreme Court in columns (1) and (2), a dummy variable identifying if a case has been reversed by the Supreme Court in columns (3) and (4), and the number of citations that a case has received in columns (5) and (6). *Judges' sentiment* is the text-based sentiment of each case. OLS are ordinary least square estimates. 2SLS estimates use as instrument the predicted case sentiment from an elastic net model based on judges' characteristics, applying cross-validation. All estimates include year circuit, circuit × year, target and legal issue fixed effects, and a dummy indicating whether the initial circuit decision was affirm or reverse. The outcome variable and main regressor are centred and standardized by target. Observations are at the case–target level. Standard errors clustered by circuit–year in parentheses.
*, **, *** indicate $p < 0.1$, $p < 0.05$, $p < 0.01$, respectively.

Specifically, all estimates include circuit × year, target and legal topic fixed effects, and a dummy indicating the direction of the Circuit Court decision (affirm or reverse).

In columns (1) and (2) of Table 1, we provide the results when using as outcome a dummy variable indicating whether a case has been reviewed by the Supreme Court. The effect is positive and statistically significant when using either an OLS or 2SLS regression. The OLS estimate suggests that a one standard deviation increase in the positivity of judicial sentiment increases the chance of a case being reviewed by the Supreme Court by 0.2%. In the 2SLS estimate the coefficient is ten times larger, therefore a one standard deviation increase in judicial sentiment would increase by 2% the probability of a Supreme Court intervention.

In columns (3) and (4) of Table 1, we use as dependent variable a dummy identifying if a case is reversed by the Supreme Court. Also in this case, the effect is positive and statistically significant in both estimates. Similarly to the results just discussed, the coefficient from 2SLS is nearly ten times larger compared to the OLS coefficient. In particular, a one standard deviation increase in our treatment will increase the chance of a case being reversed by the Supreme Court by 0.1% if estimated using OLS, and 1.1% when using 2SLS.

Finally, in columns (5) and (6) of Table 1, we focus on the effect of positive sentiment on the number of citations that a case later receives. We find again a positive and significant coefficient, which is larger in 2SLS compared to OLS. When using OLS the coefficient is 0.111, which is comparable to 0.04% of a standard deviation of the dependent variable, while in 2SLS the coefficient is 1.326, which is comparable to 0.9% of a standard deviation deviation of the dependent variable.

Thus exogenously higher sentiment in a case increases the likelihood of appeal and reversal. Yet it also increases the number of citations. One interpretation of this result is that higher sentiment makes a ruling more expressive. Hence it could attract both negative attention (as indicated by appeal) but also be more memorable or persuasive for future judges (as indicated by citations).

This application might face identification issues, in particular because of a potential violation of the exclusion restriction. We cannot rule out that judge characteristics could impact higher court decisions and citations through channels other than the expressed sentiment—for instance, via a contemporaneous effect on the actual judicial decisions that we do not observe. We can partially account for this by including as control whether the Circuit Court decision was to reverse the lower court verdict. The results are robust to including that as a control, suggesting that our effects are due to the text sentiment and not the confounded direction of the decision.

Across the different outcomes, the 2SLS coefficient is much larger in magnitude than that from OLS. There could be many reasons for this, starting with the fact that 2SLS captures a local average treatment effect. The complier cases in this setting are those where there is scope for judges to shift the sentiment—that is, legal contexts where sentiment is driven by how judges decide to frame issues, and not the issues themselves. Further, OLS could be negatively biased. For example, the measured sentiment could be lower in topics, such as death penalty cases, that are also unlikely to be reviewed or cited.

VI. Conclusion

In summary, this paper has combined natural language processing, machine learning and causal inference techniques to provide a method for analysing the impacts of judicial sentiments. There are many research opportunities opened up by these methods. Our

approach could be used to develop sentiment metrics in other corpora, such as political speeches, news articles or corporate earnings calls. One could measure sentiment toward other targets—and not just social groups, but also concepts such as democracy or inequality, for example. The cross-validated instruments approach could be applied in other circumstances with many weak instruments that are predictive of treatment. Random assignment of judges, along with judicial texts, could be used to analyse causal impacts of other features of legal language.

Our descriptive evidence shows that for most of the target groups, sentiment has been stable over time. Also, when comparing across circuits, we find that the direction of sentiment is largely the same, while the intensity might differ. This is also true when estimating differences in sentiment of judging panels with different demographic characteristics. The absence of systematic differences in aggregate sentiment is surprising, but does not mean that sentiment at the level of the decision is irrelevant. In particular, we show that within-circuit changes in judge sentiment over time are negatively related to the sentiments expressed in surveys by residents in those circuits, consistent with a backlash effect that has been discovered in other work (Chen *et al*. 2016; Wheaton 2020).

In our empirical analysis of the causal effect of judicial sentiment, we study the impacts of decision sentiment on the development of the law. We find that judge writing sentiment does have an impact on Supreme Court decisions and the number of citations. The more positive (rather than negative) the sentiment expressed in the rulings, the higher the chances that the Supreme Court will review and reverse previous decisions. Moreover, cases with more positive sentiment receive more citations. Hence we identify a dimension in judicial language (sentiment) that amplifies the influence of a judge's decision via citations, yet also increases the likelihood that the case is reviewed for error. It could be that sentiment makes the ruling more expressive, increasing both negative and positive attention to it by other judges. These results add to the literature on judicial decision-making and judicial quality (Posner 2010; Ash *et al*. 2020; Ash and MacLeod 2021).

ACKNOWLEDGMENTS

We thank Noam Yuchtman (the editor) and one anonymous referee for their very useful comments. We also thank Sherman Aline, David Cai and Léo Picard for their helpful research assistance. Open access funding provided by Eidgenossische Technische Hochschule Zurich.

NOTES

1. The database allows us to distinguish between nine categories: Criminal, Civil Rights, First Amendment, Due Process, Privacy, Labor Relations, Economic Regulation and Miscellaneous.
2. See www.cas.sc.edu/poli/juri/attributes.htm (accessed 19 October 2021).
3. A classic example shows that using the vector representation of 'king', 'man' and 'woman', the embedding model would know that the analogy of 'king' would be 'queen' via the following vector algebra: king−man+woman=queen.
4. Spirling and Rodriguez (2021) show that empirical applications are not sensitive to these default parameters (dimension and window size), so changing them should not matter much.
5. The ANES is a survey conducted every two years since 1948. It provides information about citizens' voting behaviour, as well as their attitudes.
6. The cosine similarity between two vectors is $s(\vec{v}, \vec{w}) = \vec{v} \cdot \vec{w}/(\|\vec{v}\|\|\vec{w}\|)$, which is equal to 1 minus the cosine of the angle between the vectors.
7. These values are selected via tenfold cross-validation in each of the five folds of the elastic net.
8. In unreported estimates we reach similar results using both LASSO (L1 but not L2 penalty) and ridge regression (L2 but not L1 penalty) to form the predictions (Belloni *et al*. 2012; Zou and Hastie 2005).

REFERENCES

ANGRIST, J. D. and KRUEGER, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, **13**(2), 225–35.

ASH, E. and CHEN, D. L. (2017). Judge embeddings: toward vector representations of legal belief. Technical Report.

ASH, E. and CHEN, D. L. (2019). Case vectors: spatial representations of the law using document embeddings. In M. Livermore and D. Rockmore (eds), *Law as Data*. Santa Fe, NM: Santa Fe Institute Press.

ASH, E., CHEN, D. L. and NAIDU, S. (2020). Ideas have consequences: the impact of law and economics on American justice. Center for Law & Economics Working Paper no. 4.

ASH, E. and MACLEOD, W. B. (2021). Reducing partisanship in judicial elections can improve judge quality: evidence from US state supreme courts. *Journal of Public Economics*, **201**; available online at https://doi.org/10.1016/j.jpubeco.2021.104478 (accessed 19 October 2021).

ASH, E., MORELLI, M. and VAN WEELDEN, R. (2017). Elections and divisiveness: theory and evidence. *Journal of Politics*, **79**(4), 1268–85.

BANDIERA, O., PRAT, A., HANSEN, S. and SADUN, R. (2020). CEO behavior and firm performance. *Journal of Political Economy*, **128**(4), 1325–69.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**(6), 2369–429.

BROMLEY, R. V. (1994). Journalists assess computers' value in covering US courts of appeals. *Newspaper Research Journal*, **15**(1), 2–13.

CALISKAN, A., BRYSON, J. J. and NARAYANAN, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–6.

CHEN, D. L., LEVONYAN, V. and YEH, S. (2016). Policies affect preferences: evidence from random variation in abortion jurisprudence. Toulouse School of Economics Working Paper no. 16-723.

CHEN, J., CHEN, D. L. and LEWIS, G. (2020). Mostly harmless machine learning: learning optimal instruments in linear IV models. arXiv preprint arXiv:2011.06158.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, **107**(5), 261–5.

CLARK, T. S., STATON, J. K., WANG, Y. and AGICHTEIN, E. (2018). Using twitter to study public discourse in the wake of judicial decisions: public reactions to the Supreme Court's same-sex-marriage cases. *Journal of Law and Courts*, **6**(1), 93–126.

DAI, A. M., OLAH, C. and LE, Q. V. (2015). Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

DI TELLA, R. and SCHARGRODSKY, E. (2013). Criminal recidivism after prison and electronic monitoring. *Journal of Political Economy*, **121**(1), 28–73.

DOBBIE, W. and SONG, J. (2015). Debt relief and debtor outcomes: measuring the effects of consumer bankruptcy protection. *American Economic Review*, **105**(3), 1272–311.

DRACA, M. and SCHWARZ, C. (2019). How polarized are citizens?Measuring ideology from the ground-up. Working Paper, University of Warwick.

GALASSO, A. and SCHANKERMAN, M. (2014). Patents and cumulative innovation: causal evidence from the courts. *Quarterly Journal of Economics*, **130**(1), 317–69.

GENTZKOW, M. and SHAPIRO, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica*, **78**(1), 35–71.

GENTZKOW, M., SHAPIRO, J. M. and TADDY, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, **87**(4), 1307–40.

KLING, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, **96**(3), 863–76.

KOZLOWSKI, A. C., TADDY, M. and EVANS, J. A. (2019). The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review*, **84**(5), 905–49.

LE, Q. V. and MIKOLOV, T. (2014). Distributed representations of sentences and documents. CoRR abs/1405.4053, –.

LIM, C. S., SNYDER, J. M. Jr and STRÖMBERG, D. (2015). The judge, the politician, and the press: newspaper coverage and criminal sentencing across electoral systems. *American Economic Journal: Applied Economics*, **7**(4), 103–35.

MAESTAS, N., MULLEN, K. J. and STRAND, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review*, **103**(5), 1797–829.

MIKOLOV, T., CHEN, K., CORRADO, G. and DEAN, J. (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781, –.
POSNER, R. A. (2010). *How Judges Think*. Cambridge, MA: Harvard University Press.
SAMPAT, B. and WILLIAMS, H. L. (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, **109**(1), 203–36.
SPIRLING, A. and RODRIGUEZ, P. (2021). Word embeddings: what works, what doesn't, and how to tell the difference for applied research. *Journal of Politics*, forthcoming.
WEINRIB, L. M. (2012). The sex side of civil liberties: United States v. Dennett and the changing face of free speech. *Law and History Review*, **30**(2), 325–86.
WHEATON, B. (2020). Laws, beliefs, and backlash. Unpublished Working Paper.
ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–20.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

A.1 List of words to identify the attributes and target groups
A.2 Other Figures and Tables