Measuring Judicial Sentiment: Methods and Application to US Circuit Courts

Elliott Ash, Daniel L. Chen, Sergio Galletta

Measuring Judicial Sentiment: Methods and Application to U.S. Circuit Courts *

Sergio Galletta[1], Elliott Ash[2], Daniel L. Chen[3]

[1] *University of Bergamo*
[2] *ETH Zürich*
[3] *Toulouse Institute for Advanced Study*

Abstract

This paper provides a general method for analyzing the sentiments expressed in the language of judicial rulings. We apply natural language processing tools to the text of U.S. appellate court opinions to extrapolate judges' sentiments (positive/good vs. negative/bad) toward 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.

*We thank Sherman Aline, David Cai and Léo Picard for their helpful research assistance.

Email addresses: sergio.galletta@unibg.it (Sergio Galletta), ashe@ethz.ch (Elliott Ash), dlchen@nber.org (Daniel L. Chen)

1. Introduction

Law is composed of natural language, and 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 analyzing 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 U.S. Circuit Court opinions. In particular, we draw upon 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 analyze 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 straight-forward 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). 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 sentiment 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 to be 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 behaviors (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 analyzing millions of books published over 100 years.

Methodologically, our approach to estimate causal effects is close to 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 2 describes the insti-

tutional background and data. Section 3 describes the method use to measure judges' sentiment. Section 4 details the instrumental variables approach. Section 5 provides descriptive evidence about the variation in expressed sentiment, while Section 6 reports an application of our methodology. Section 7 concludes.

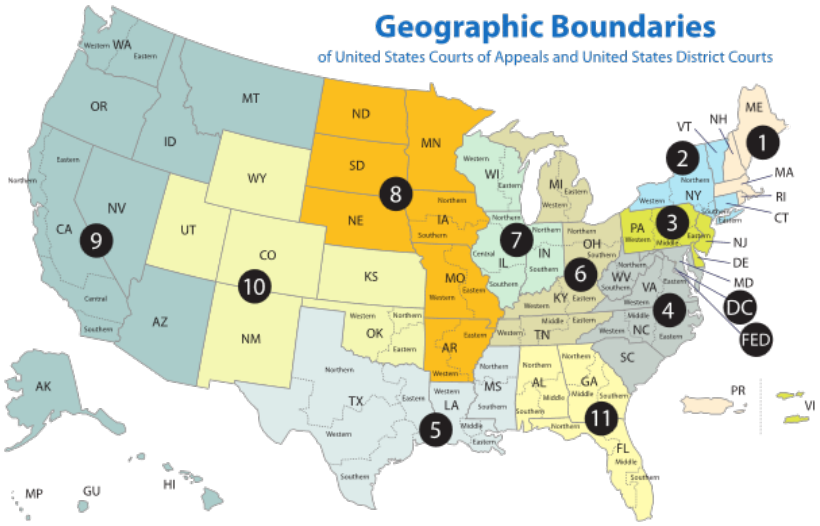
2. Institutional Setting and Data

2.1. The U.S. federal court system

The U.S. Federal Courts system is organized on three levels: the national level (Supreme Court), intermediate level (Circuit Courts) and local level (District Courts). Our contribution exploit features that are specific of the intermediate level. The Circuit Courts play a crucial role as their judges decide whether the decisions taken by the District Court were erroneous.

There are 12 regional U.S. Circuit Courts. Each of these courts is responsible for 3-9 states (see Figure 1), a part from to the U.S. Circuit court of District the of Columbia.[1] For each case there are assigned three life-tenure judges. On average a Circuit has 17 judges, with a minimum of 8 and maximum of 40.

Figure 1: U.S. Courts of Appeals



[1]There is also a the 13th court of appeals (Federal Circuit) which has nationwide jurisdiction on specific subjects.

Circuit court judges are powerful forces in U.S. politics and culture. 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 elites in government and media.

Evidence of elite response 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 favor of stronger free-speech protections. Clark et al. (2018) find significant responses on Twitter after several court decisions. Bromley (1994) is an early paper documenting how journalists do research on circuit court opinions. Lim et al. (2015) document the frequent coverage of criminal decisions in newspapers.

2.2. Data

We have assembled data from a range of sources. To create the judges' sentiment measure, we use the complete collection of United States Courts of Appeals opinions from 1961 to 2013. The corpus includes all published cases and comes from Bloomberg Law. For each case we also use additional information such as whether the Supreme Court reviewed the case, if the Supreme Court reversed the decision, the number of citations, if the case was reversed by the Circuit Court, and general category labels.[2] We use these data in the empirical application section.

Further, we collected the biographical information of judges that have been assigned to at least one case in the same period. We match each judge with data from the Federal Appeals and District Court Attribute Data.[3] 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 affiliations.

[2]The database allows to distinguish between 9 categories: Criminal, Civil Rights, First Amendment, Due Process, Privacy, Labor Relations, Economic Regulation and Miscellaneous.

[3]http://www.cas.sc.edu/poli/juri/attributes.htm

3. A Measure of Judicial Sentiment

To measure judicial sentiment, we apply a text embedding model to the text of U.S. Circuit Court opinions. Embedding models are a recent NLP techniques that have been mainly implemented in computational linguistics for prediction tasks 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 assign 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 is the similarity of the context. Moreover, trained embeddings encode meaningful information about analogies.[4]

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., democrats, republicans, business, etc.). This idea is closely related to Caliskan et al. (2017) who used word space similarly to gauge biased associations in text. The potential of this approach is also explored in a recent paper by Kozlowski et al. (2019). By using similar tools 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, to create our measure of judicial sentiment we use the collection of opinions of all published cases of United States Courts of Appeals. 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 suggested, words that tend to have similar

[4] 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.

contexts are located near each other (we used window size of five).[5] 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 analyze similarities and analogical relations between documents.

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).[6] 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 Appendix section A.1). Blacks, for example, are identified off of black, blacks, african, african, african-american, african-americans, negro, and negroes. In Appendix Figure A.1, we provide word clouds that report the words most associated with each target.

In the caselaw corpus, we compute the cosine similarity of each sentence vector to each of the targets.[7] 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 $W_i^k = \frac{1}{|D_i|} \sum_{i \in D_i} W_{id}^k$, where D_i is the total number 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", "favorable", "good") and negative words (e.g., "cold", "unfavorable", "bad") (see Appendix section A.1). Figure 2 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} 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

[5] Spirling and Rodriguez (2019) show that these default parameters (dimension and window size) tend to work well and changing them should not matter much in our type of empirical application.

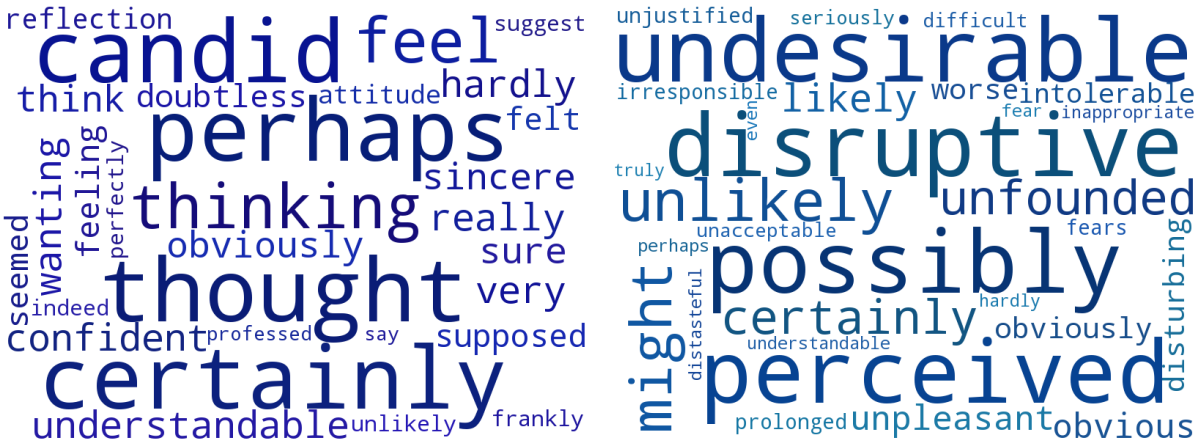
[6] ANES is a survey conducted every two years since 1948 and provides information about citizens voting behavior, as well as their attitudes.

[7] The cosine similarity between two vectors is $s(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\| \|\vec{w}\|}$, which is equal to one minus the cosine of the angle between the vectors.

Figure 2: Positive and Negative Sentiment Language

(a) Positive Sentiment

(b) Negative Sentiment



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.

case-level sentiment towards target k as $S_{ik} = \sum_{i \in D_i} S_{id} \cdot W_{id}^k$, the dot product of these two vectors. These information can be also summarized at a higher level of aggregation by computing the average sentiment of the cases that occurred at the level of interest. 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 toward k for cases in circuit-year ct.

4. Estimating the Effect of Judicial Sentiment

In this section we describe a generic framework to use to study the causal effects of judge sentiment in the context of the United States Courts of Appeals. Very often to address research questions related to the effect of judge sentiment a simple OLS would not be sufficient to provide causal evidence as the regression would have endogeneity issues and resulting estimates would likely be biased.

To account for these endogeneity concerns we suggest an instrumental variables strategy which exploits the random assignment of judges to federal circuit courts as a source of exogenous variation. In particular, we take advantage of the evidence that judge characteristics are good cross-validated predictors of expressed sentiments, together with the fact that the personal characteristics of the 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 in the related literature 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 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 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.

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), therefore 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). Specifically, to predict sentiment using the judges' characteristics we can use different regularization methods such as LASSO, ridge regression or elastic net. Next, we run the cross-validated prediction. The predicted endogenous regressor is the instrument in our two-stage least-squares 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 toward 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:

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

where the α's 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.

5. Descriptive Evidence on Judge Sentiment

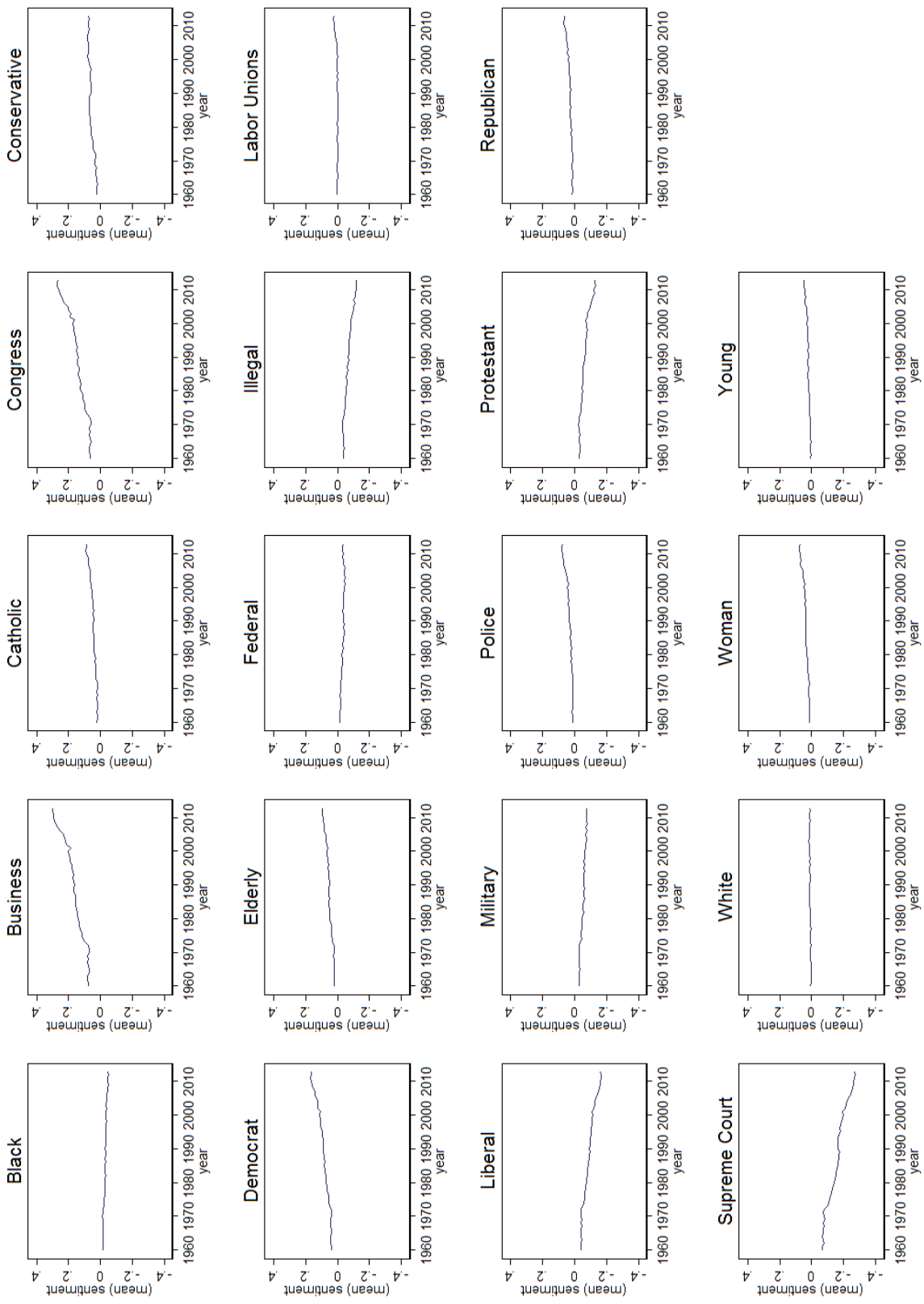
In this section we investigate the details of our measure of judicial sentiment, looking at its variation across different dimensions.

We begin by looking at variation of expressed sentiment over time. In Figure 3 we display the results when focusing on time variation by showing 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 toward business, Catholics, and Democrats. Meanwhile, there is a negative trend for liberal and supreme court. The increase in positive sentiment toward business could be part of a previously noted increasing economic conservatism in the judiciary (Ash et al., 2020).

Next, in Figure 4, we show 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. Only the D.C. Circuit (indicated as the 12th district in the Figure) reports sentiment that differ to the ones from the other circuits, when the target groups are labour union and federal government. 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.

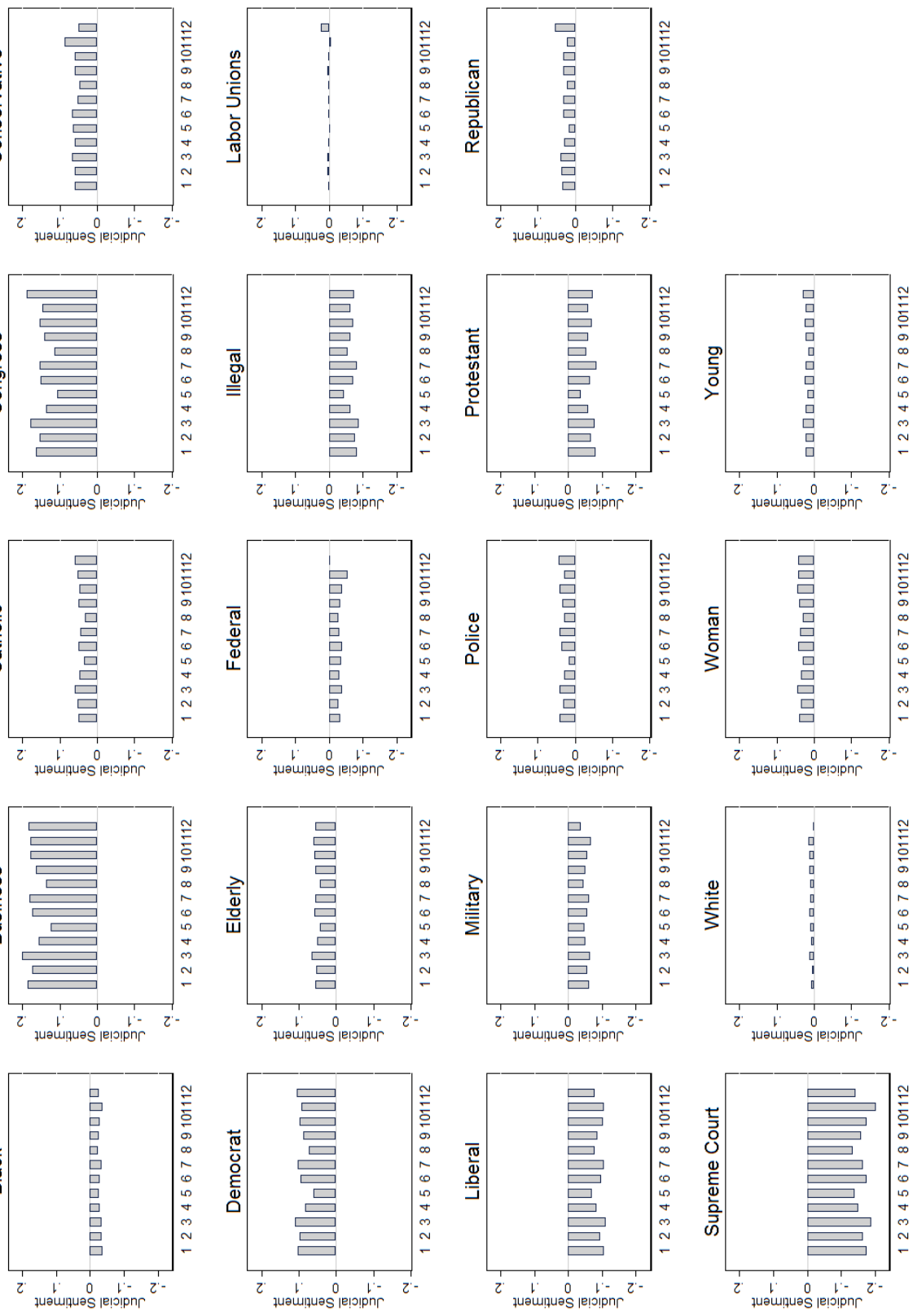
Further evidence about the variation in sentiment can be illustrated by exploiting the individual characteristics of the three judges assigned to the cases. For this purposes we display the mean level of sentiment for each target group reported by different composition of the judges' panel. We display this results in Figure 5 in four different panels. In panel (a), we show the mean sentiment towards different target groups depending on whether the judges were appointed by a Democratic or a Republican president. In panel (b) we report results about the gender composition of the group of judges. In this case we compare pool of judges composed by all male with the ones that has at least one female. In panel (c) we exploit difference in race comparing panels that are composed by all white judges with panels that have 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

Figure 3: Judicial Sentiment, over time



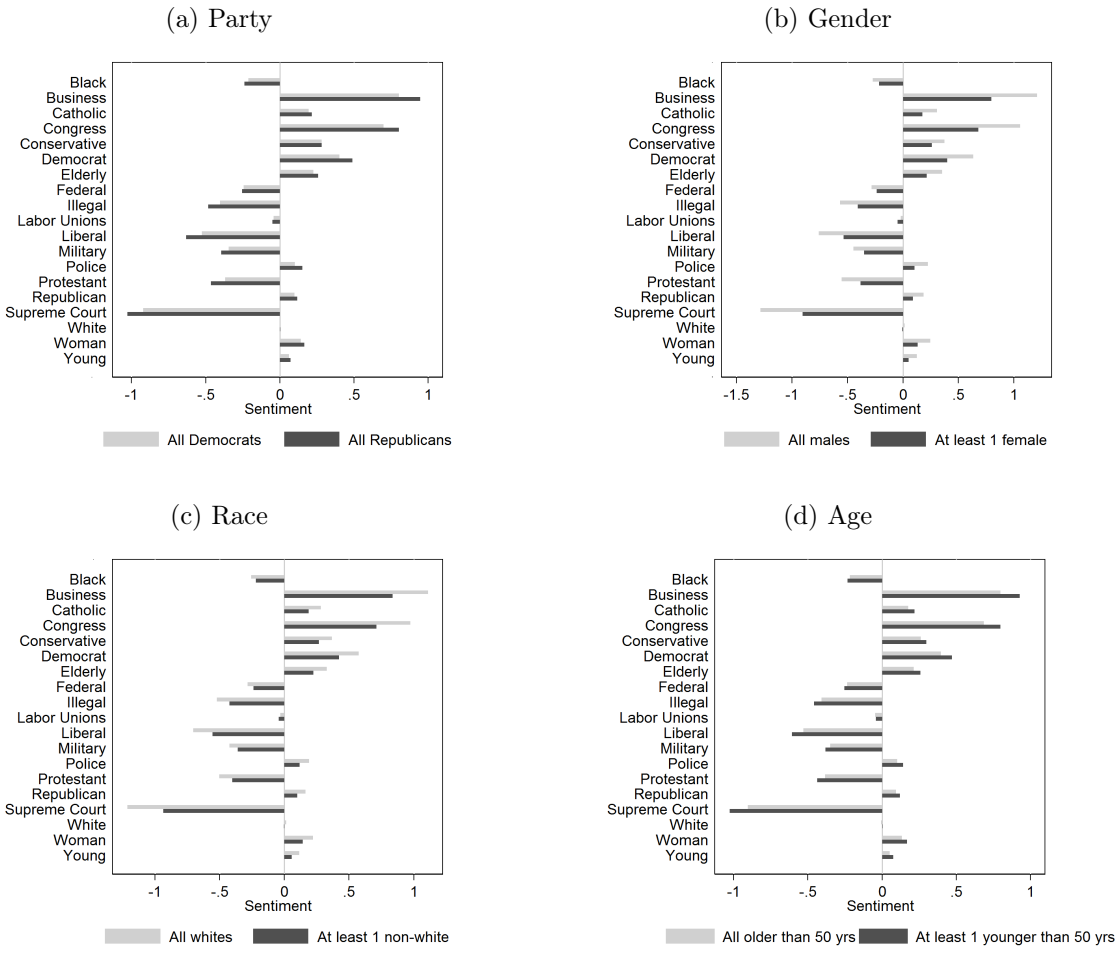
Notes: Judicial sentiment variation across years.

Figure 4: Judicial Sentiment, by circuit



Notes: Judicial sentiment variation across circuit courts.

Figure 5: Demographic characteristics of the pool of judges



Notes: Judicial sentiment variation depending on the demographic characteristics of the assigned panel of judges.

members are all older 50 years and those that have at least one judge that is younger than 50 years. For each of these items, we found qualitatively similar relationships when residualizing the variables on court-year fixed effects beforehand.

The overall picture suggests that there are differences in sentiment depending on the demographic composition of the responsible judges. Yet, these difference are not that strong in the sense that none of the comparisons are there sentiments that have an opposite sign. Some interesting patterns, for instance, are that when the target is Black we find that the sentiment is generally negative, and relatively larger when the judges are all republicans, all males, all white and all older than 50 years. When the target is Business there is generally a positive sentiment, and this is higher when all judges are republicans, all males, all white, and at least one judges is younger than 50 years.

Not all comparisons seem to provide the expected results. In particular, we find that when all judges are appointed by a Republican president the sentiment is higher towards the target group Democrat, compared to the sentiment expressed by judges selected by a Democrat president. Also when looking at the gender composition of the panel we see that the sentiment towards women is higher when the pool of judges is formed only by males.

These results could be seen as an example of the potential limits that using text analysis can have in catching nuances in the rulings, and therefore a call for researchers to be cautious in the interpretation of the evidence.

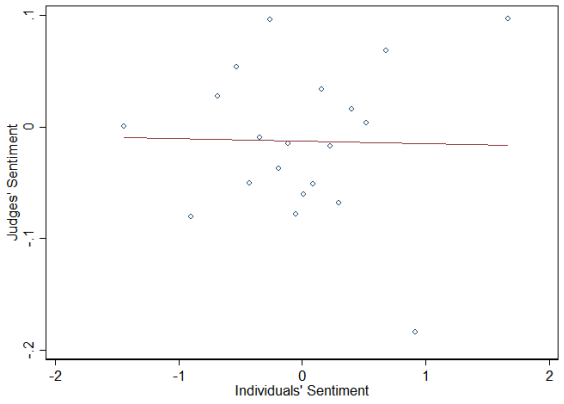
As a final descriptive exercise we ask whether the sentiments expressed by judges in opinions are correlated with local circuit residents' sentiments reported in surveys. For this analysis we measure individuals' preferences towards the set of 19 target groups with information from the ANES. The feeling thermometer questions ask about attitudes towards a specific target group by choosing a value from 0 to 100 (see Appendix Figure A.2). A value closer to 100 reveals that the respondent feels warmly or favorably towards the target group, while a closer to 0 means cold or unfavorable feelings towards the target group.

Figure 6 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

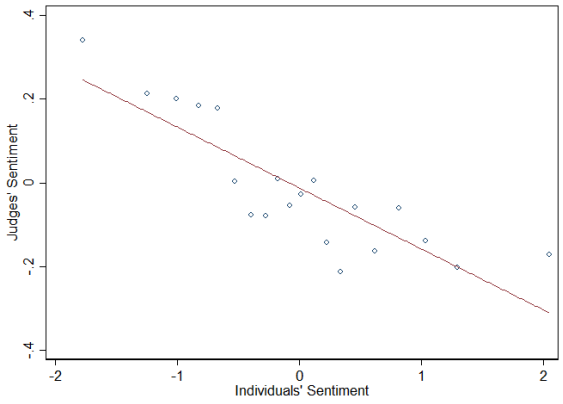
over time. These statistics provides some additional descriptive evidence to complement the causal evidence on abortion attitudes provided by Chen et al. (2016).

Figure 6: Correlation between judges' and individuals' sentiment

(a) Across circuits



(b) Within circuits



Notes: Binscatter diagrams displaying the correlation between judges' sentiment and individuals' sentiment. Panel A includes as controls year fixed effects × target fixed effects. Panel B includes as controls circuit fixed effects × target fixed effects.

6. Empirical application

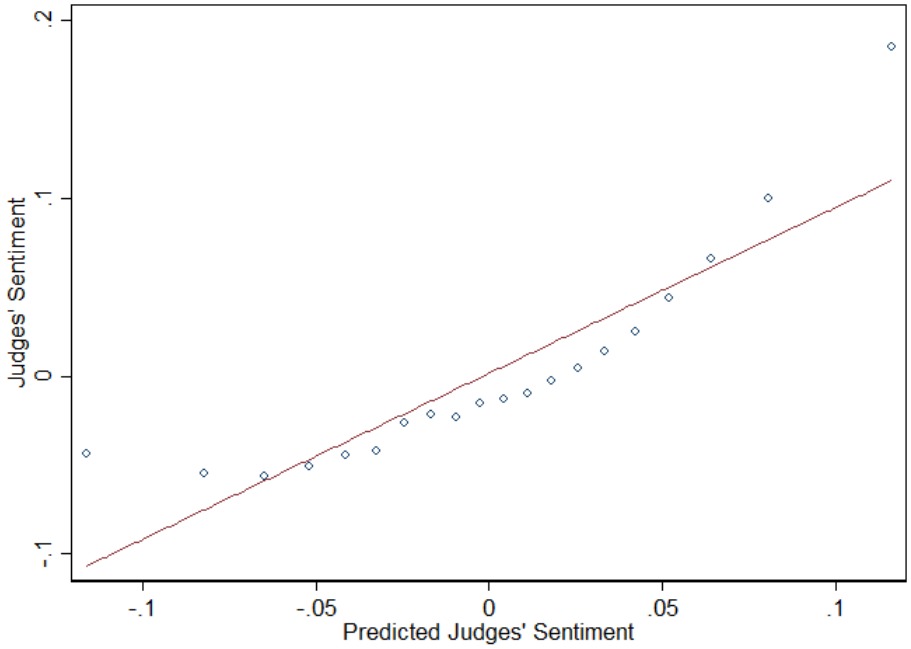
In this section we provide an example of the implementation of our methodology by studying the relationship between judges' sentiment expressed in their rulings and the influence of their decisions into the law. This is relevant in a common-law system where judicial rulings are binding precedent on lower courts. They also speak to the linguistic factors that judges find persuasive.

Specifically, we study the effect of expressed sentiment on variables that are good proxies for legal impact. These include the likelihood that the Supreme Court reviews a circuit court decision, the likelihood 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 to construct the instrumental variable from judges' characteristics as described in Section 4. To prepare the data for the prediction task from which the instrument originates, we standardize to variance one 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, 2005a). The predictions are then formed using a five-fold cross-validation. We learned the cost-minimizing penalties: L1 = 0.2 and L2 = 0.8 and a general penalty $\lambda = 0.00013$.[8] 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.[9]

Figure 7: First stage relationship



Notes: Binscatter diagram for the first stage relationship (Coeff.= 0.931, st. err.= 0.056, R^2 = 0.006)

Next, we implement the first stage equation 2 and we confirm that the instrument is strongly predictive of sentiment (coeff= 0.931, st. err.= 0.056 and therefore F-stat= 278), but far from being collinear ($R^2 = 0.006$), as also shown in Figure 7's scatter plot.

Finally, in Table 1 we present our main findings reporting the estimate of the effect of judicial sentiment on different outcomes. We report both OLS estimates and results based on the suggested instrumental variables approach, always including a relevant set of fixed effects and controls.[10] Specifically, all estimates include year FE, circuit FE,

[8] These values are selected via 10-fold cross-validation in each of the five estimates of the elastic net.

[9] 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, 2005b).

[10] We find similar results in terms of statistical significance also when selecting one-tenth of our sample.

Table 1: Empirical Application

	Supreme Court Reviewed		Supreme Court Reversed		Number of Citations	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Positive 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-stat		278		278		278
N 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 (1-2), a dummy variable identifying if a case has been reversed by the Supreme Court in columns (3-4) and the number of citations that a case has received in column (5-6). *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 based on judges characteristics, applying cross-validation. All estimates include year FE, circuit FE, circuit × year FE, target FE, geniss FE and a dummy indicating whether the verdict was reversed by circuit judges. The dependent and the main regressor are centered and standardized by target. Observations are at the case-target level. Standard errors clustered by circuit-year in parenthesis. * $p < 0.1$, ** $p < 0.05$ and *** $p < 0.01$.

circuit × year FE, target FE, legal topic FE, and a dummy indicating whether the verdict was reversed in appeal by circuit judges.

In columns (1-2) we provide the results when using as outcome variable a dummy identifying whether a case has been reviewed by the Supreme Court. The effect is positive and statistically significant when using either an OLS or the 2SLS regression. The OLS estimate suggests that a one standard deviation increase in the positivity of judicial sentiment increases the chance of a case to be reviewed by the Supreme court by 0.2%. In the 2SLS the coefficient is 10-times larger, and therefore a one standard deviation increase in judicial sentiment would increase by 2% the probability of a Supreme Court intervention.

In columns (3-4) we use as a dependent variable a dummy identifying if a case has been 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 the 2SLS is nearly 10-times larger compared to the OLS coefficient. In particular, a one standard deviation increase in our treatment will increase the chance of a case to be reversed by the Supreme court by 0.1% if estimated using the OLS and 1.1% when using the 2SLS.

Finally, in columns (5-6) we focused on the effect of positive sentiment on number of citations that a case later receives. We find again a positive and significant coefficient, which is quite larger in the 2SLS compared to the OLS. When using the OLS the coef-

ficient is 0.111, which is comparable to 0.04% of a standard deviation of the dependent variable. While in the 2SLS the coefficient is 1.326, which is comparable to 0.9% of a standard deviation deviation of the dependent variable.

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 other channels 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 issue 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 our effects are due to the text sentiment and not the confounded direction of the decision.

7. Conclusion

In summary, this paper has combined natural language processing, machine learning, and causal inference techniques to provide a method for analyzing 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 or news articles, and toward other targets (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 analyze causal impacts of other features of legal language.

Descriptive evidence show 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.

In the empirical application, we study the impacts of judge writing 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 the sentiment (rather than negative) expressed in the rulings, the higher are the chances that the Supreme Court will review and reverse previous decisions. Moreover, cases with more positive sentiment receive more citations. These are relevant results that add to the literature on the determinants of judicial decision-making (e.g Ash et al., 2020).

Bibliography

References

Angrist, J. D. and A. B. Krueger (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics 13*(2), 225–235.

Ash, E. and D. L. Chen (2017). Judge embeddings: Toward vector representations of legal belief. Technical report.

Ash, E. and D. L. Chen (2019). Case vectors: Spatial representations of the law using document embeddings. *Law as Data, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore*.

Ash, E., D. L. Chen, and S. Naidu (2020). Ideas have consequences: the impact of law and economics on american justice. *Center for Law & Economics Working Paper Series 4*.

Ash, E., M. Morelli, and R. Van Weelden (2017). Elections and divisiveness: Theory and evidence. *The Journal of Politics 79*(4), 1268–1285.

Bandiera, O., A. Prat, S. Hansen, and R. Sadun (2020). Ceo behavior and firm performance. *Journal of Political Economy 0*(0), 000–000.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

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

Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science 356*(6334), 183–186.

Chen, D. L., V. Levonyan, and S. Yeh (2016, October). Policies Affect Preferences: Evidence from Random Variation in Abortion Jurisprudence. TSE Working Papers 16-723, Toulouse School of Economics (TSE).

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review 107*(5), 261–65.

Clark, T. S., J. K. Staton, Y. Wang, and E. Agichtein (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., C. Olah, and Q. V. Le (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

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

Dobbie, W. and J. Song (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review 105*(3), 1272–1311.

Draca, M. and C. Schwarz (2019). How polarized are citizens? measuring ideology from the ground-up. Working papers, University of Warwick.

Galasso, A. and M. Schankerman (2014). Patents and cumulative innovation: Causal evidence from the courts. *The Quarterly Journal of Economics 130*(1), 317–369.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from us daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review 96*(3), 863–876.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review 84*(5), 905–949.

Le, Q. V. and T. Mikolov (2014). Distributed representations of sentences and documents. *CoRR abs/1405.4053*, –.

Lim, C. S., J. M. Snyder Jr, and D. Strömberg (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., K. J. Mullen, and A. Strand (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review 103*(5), 1797–1829.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*, –.

Sampat, B. and H. L. Williams (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review 109*(1), 203–36.

Spirling, A. and P. Rodriguez (2019). Word embeddings: What works, what doesn't, and how to tell the difference for applied research.

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), 325–386.

Zou, H. and T. Hastie (2005a). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Zou, H. and T. Hastie (2005b). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B 67*, 301–320.

A. Appendix

A.1. List of words to identify the attributes and target groups

Attributes

Negative: cold, unfavorable, bad, adverse, antagonistic, calamitous, damaging, destructive, disadvantageous, hostile, negative, objectionable, ominous, troublesome, unfriendly, contrary, discommodious, ill, ill-advised, improper, inadvisable, inauspicious, inconvenient, inexpedient, infelicitous, inimical, inopportune, late, low, malapropos, opposed, poor, regrettable, tardy, threatening, unfit, unfortunate, unlucky, unpromising, unpropitious, unseasonable, unseemly, unsuited, untimely, untoward, wrong.

Positive: warm, favorable, good, agreeable, benign, encouraging, positive, supportive, sympathetic, acclamatory, affirmative, amicable, approbative, approbatory, assenting, benevolent, benignant, commending, complimentary, enthusiastic, inclined, kind, kindly, laudatory, okay, praiseful, predisposed, reassuring, recommendatory, understanding, welcoming, well-disposed, well-intentioned.

Targets

Black: blacks, black, african, african-american, african-americans, negro, negroes
Business: business, businesses, corporation, corporations, factory, firm, market, organization, partnership, shop, store, venture
Catholic: catholics, catholic
Congress: congress, parliament, legislature, senate, house, representative, senators, representatives
Conservative: conservatives, conservative
Democrat: democrat, democrats
Elderly: elderly, aged, old
Federal government: federal, government, executive
Illegal immigrants: illegal, immigrants, undocumented
Labor unions: labor, unions, union, trade-union
Liberal: liberals, liberal
Military: military, army
Police: policemen, police, policeman

Protestant: protestant, protestants

Republican: republican, republicans

Supreme Court: supreme, court

White: whites, white, caucasian, caucasians

Woman: woman, women

Young: youngster, youth, budding, adolescent

A.2. Other Figures and Tables

Table A.1: Summary statistics

Variable	Mean	Std. Dev.	Min	Max	N
Supreme Court reviewed	0.015	0.121	0	1	3,377,250
Supreme Court reversed	0.005	0.068	0	1	3,377,250
Number of citations	9.338	24.795	0	9193	3,377,250
Judges' sentiment	0.002	0.994	-40.968	82.232	3,377,250
Predicted judges' sentiment (instrument)	0	0.071	-0.417	0.439	3,377,250
Circuit Court reversed	0.175	0.38	0	1	3,377,250

Figure A.1: Word Clouds for Target Groups

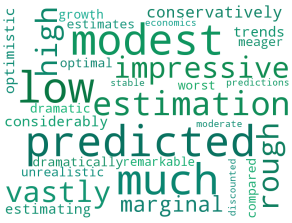
(a) Black

(b) Business

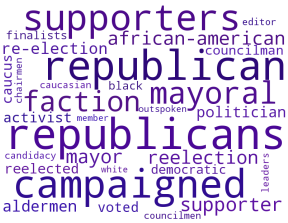
(c) Catholic

(d) Congress

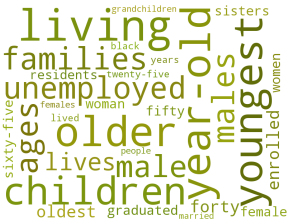
(e) Conservative



(f) Democrat



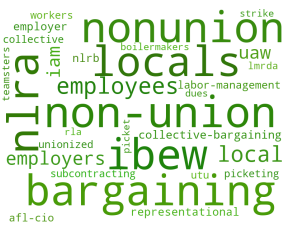
(g) Elderly



(h) Federal Gov.

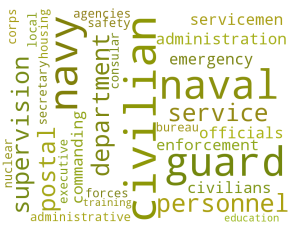
(i) Illegal Immigrants

(j) Labor Unions

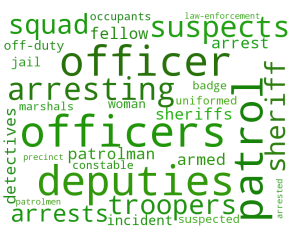


(k) Liberal

(l) Military

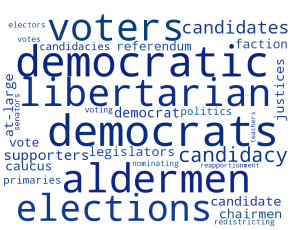


(m) Police



(n) Protestant

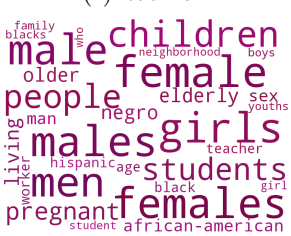
(o) Republican



(p) Supreme Court

(q) White

(r) Woman



(s) Young

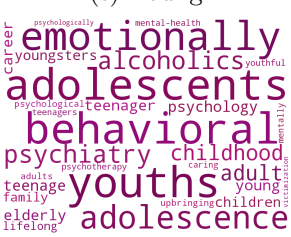


Figure A.2: Example Thermometer Question - ANES 2012