Predicting Punitiveness from Judicial Corpora

Elliott Ash and Daniel Chen

June 28, 2017

Abstract

Using over 1 million sentencing decisions linked to judge identity and the digital corpora of U.S. district court opinions, we show that judges' writings can predict average harshness and racial and sex disparities in sentencing decisions. We document significant reductions in mean square error relative to a naive prediction (the mean of the training data) on the test dataset by approximately 24 percent in predicting punitiveness.

1 Introduction

Previous work has documented large disparities in the harshness of criminal sentencing across judges for similar infractions (cite). This paper shows that these disparities can be explained and predicted by the textual features of a judge's written opinions. Moreover, the text can also predict judge-level tendencies to give disparate sentences to defendants of different races.

One interpretation is that differences in judge writing style reflect important differences in judge ideology, which lead to systematic differences in criminal-justice outcomes.

This paper builds on previous work on judicial disparities in sentencing harshness. We show that the text features of a judge's opinions have predictive power for another component of their work – the length of criminal sentences they impose.

2 Background

Federal district court judges are responsible for overseeing trials and issuing sentences for federal crimes in the United States. While the statutes specify guidelines for prison sentences, judges have significant discretion in these decisions. A recent analysis of federal court cases gathered by the Transactional Records Access Clearinghouse (TRAC) from FY 2007 to FY 2011 shows wide variation in sentencing conditioned on the crime and district [NYT02]. One study finds that after controlling for statutory, demographic, and socioeconomic variables, blacks, males, uneducated, and low-income groups tend to receive longer sentences than their peers [Mus01].

3 Data

The federal sentencing data comes from the U.S. Sentencing Commission. The publicly available dataset does not include judge information (e.g. names, demographic characteristics, etc). The full dataset has 930,000 cases for the years 1991-2012. For the years 2002 through 2011 (230,000 cases), the data include a richer set of covariates. These include the set of charges brought against the defendant, whether the case went to trial, and gender/race of the defendant. Figure 1 shows the distribution of crimes charged to the defendant.

Using FOIA requests, the dataset has been merged with judge information and published on Transactional Records Access Clearinghouse (TRAC). First, this match allows us to match the sentencing data with judge biographical information. This information was collected by the Federal Judicial Center (FJC) and supplemented by additional data collection. The judge covariates include age, gender, race, party of appointing president, religion, and ABA rating,

In addition, we have matched district court judges to their corpus of written judicial opinions. The opinions were obtained from CourtListener (courtlistener.com), an online database and search engine for legal opinions maintained by the Free Law Project. CourtListener has nearly complete coverage of the federal district courts for the years of our data set (1991-2012).

We consider various defendant and judge characteristics that may help predict

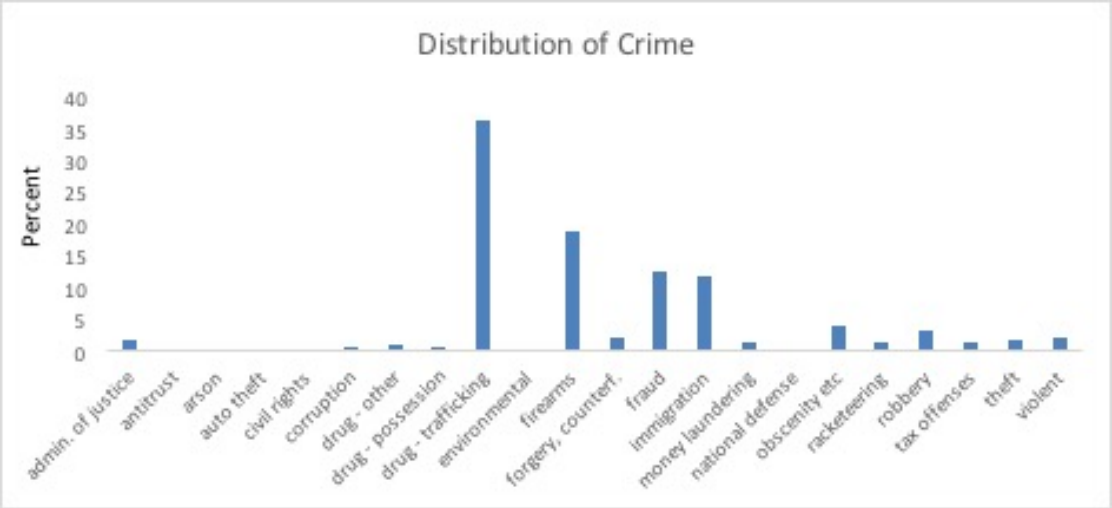


Figure 1: Distribution of Primary Crime Committed

disparity in sentencing. For judge characteristics, these are:

4 Criminal Sentencing Outcomes

This sentence describes how our outcome variables are constructed from the TRAC sentencing data.

4.1 Sentencing Harshness

We have data on case i overseen by judge j in court k at year t. We are observing sentencing decisions, after a defendant has been convicted of a crime. Given the crime in case i, the sentencing guidelines specify a minimum m_i and a maximum M_i. The judge then chooses sentence length s_i based on these guidelines. Note that this excludes any other sentencing (i.e. probation, house arrest, etc) handed down by the judge.

We define harshness H_{ijkt} following Lim et. al. (2015) [ea15]:

$$H_{ijkt} = \frac{s_i - m_i}{M_i - m_{ini}}$$

Before using this measure as an outcome in our models, we de-mean H_{ijkt} by court-year. Cases are randomly assigned to judges at the court-year level, so this purges

the outcome variable of any selection bias due to judge characteristics (conditional on court-year).

We are interested in relating the text features of a judge's written opinions at time t to this sentencing behavior at time t. Therefore we define H_{jkt} as the average (de-meaned) harshness of sentences issued by j during year t. The data include 4,800 judge-year observations.

For the larger data set that has not been merged to defendant characteristics, we do not know the crime committed and therefore lack the statuory minima and maxima for each case. For these data, we use S_{ijkt}, the log sentence length de-meaned by court-year. And again, we define S_{jkt} as the mean log sentence length for judge j during year t. For this larger data-set, we have 14,600 judge-year observations.

4.2 Sentencing Disparities by Defendant Race and Gender

An important feature of our data is that it includes some defendant demographic characteristics, namely race and gender. In this section we describe how we produce measures by judge-year for demographic disparities in sentencing.

As before, we have measures of sentence length and sentence harshness by crime, where H_{ijkt} is the harshness in case i by judge j in district k during year t, de-meaned by circuit year. We then compute the sentencing harshness of judge j separately for group g, H^g_{jkt}, where g can index race (black or white) or gender (male or female). We then compute the disparity between group g_0 and group g_1 as the ratio of or difference between the harshness levels of the the two groups.

Note that after dropping any missing values due to the possibility that the judge only saw cases of one race or one gender, our dataset shrinks to approximately 3500 judge-years for the sex disparity measure and to approximately 4200 judge-yearsfor the race disparity measure.

5 Construction of Text Data

We use the full set of judicial opinions written by federal district court judges. We adopt NLP techniques to construct a variety of text features for the prediction

task.

First, we count all words and filter out rare words. We then form n-grams up to length four. We use TF-IDF weighting to down-weight words used in all documents (and therefore less discriminative of judge style).

6 Methods

We focus on three models to analyze the predictiveness of the text features. Because we have a high-dimensional feature set, we need regularization. First, we use elastic net, which uses a combined lasso and ridge penalty to penalize the coefficients and exclude weak predictors ($\alpha = 0.1$ and $l1$ ratio $= 0.7$). Second, we use random forest regression.

For our analysis we split the dataset 80/20 for training and testing, respectively. Once we've trained our model, we compare the predicted disparity measures from the test dataset to the actual disparity measures and calculate the mean squared error. Our baseline is a "naive prediction" where for each test observation we predict the mean of the disparity measures from the training data. Any improvement from the "naive prediction" suggest the model has some predictive lift.

7 Results

7.1 Main Results

Our primary results are in tables 1 through 3.

Table 1 reports the mean squared error for out-of-sample prediction of sentencing harshness. The model without text features results in a small improvement on the naive prediction. When adding n-grams, however, we see a significant decrease in mean squared error.

For the disparity measures, we again see a performance improvement from text features (Table 2 & 3). While the prediction is only about twice as good as the naive prediction, that is a promising step that one can recover judge sentencing preferences from their writing.

Table 1: Mean Squared Error - Predicting Harshness

	Random Forest	Elastic Net
Naïve Prediction	0.0065	0.0065
No Text Features	0.0054	
Unigrams	0.0041	0.0047
Bigrams	0.0044	0.0047
Trigrams	0.0043	0.0047
Fourgrams	0.0043	0.0047
Fivegrams	0.0042	0.0047

Table 2: MSE - Disparity Measure: Sex

	Random Forest	Elastic net
Naïve Prediction	0.0199	0.0199
No Text Features	0.0168	0.0199
Unigrams	0.0177	0.0195
Bigrams	0.0167	0.0195
Trigrams	0.0159	0.0195
Fourgrams	0.0159	0.0195
Fivegrams	0.0164	0.0195

Table 3: MSE - Disparity Measure: Race

	Random Forest	Elastic net
Naïve Prediction	0.0233	0.0233
No Text Features	0.0213	
Unigrams	0.0184	0.0211
Bigrams	0.0180	0.0211
Trigrams	0.0186	0.0211
Fourgrams	0.0179	0.0211
Fivegrams	0.0181	0.0211

7.2 Inspection of Feature Importance

We try to interpret the model by inspecting the most predictive phrases. To do so we use the feature importances generated by a random forest regression. These features result in the largest change in predictive power when included or excluded. These are reported as word clouds in Figures 2, 3, and 4. Words that appear larger on the plots are features that the random forest measure deemed important in measuring the sentencing disparity variable.

Figure 2: Most predictive bigrams and trigrams for sentencing harshness

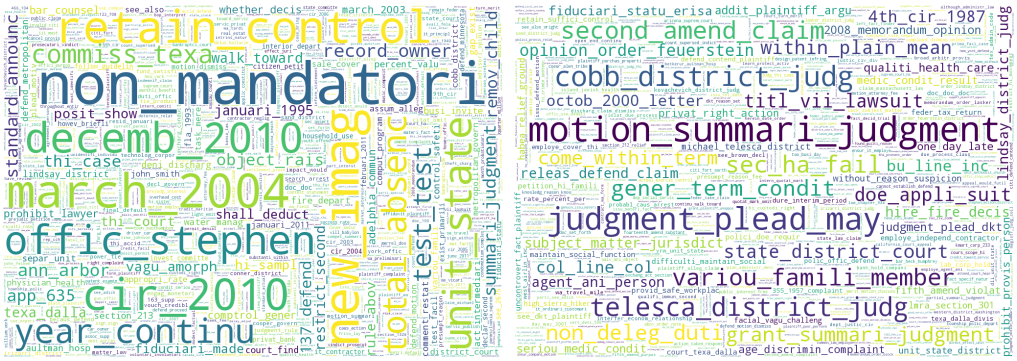
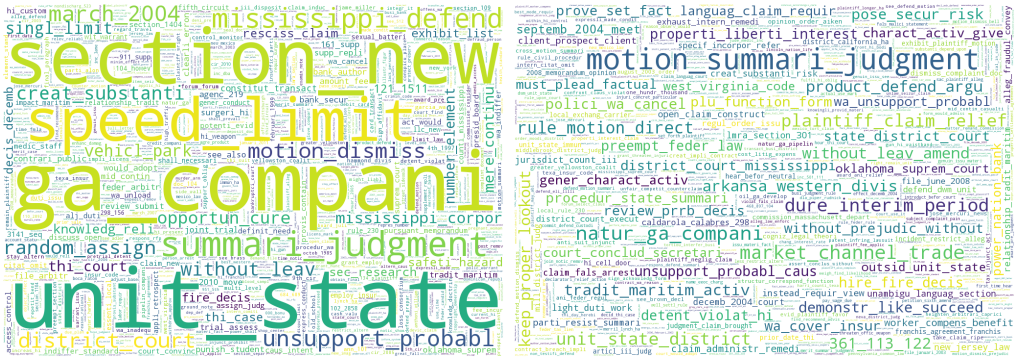
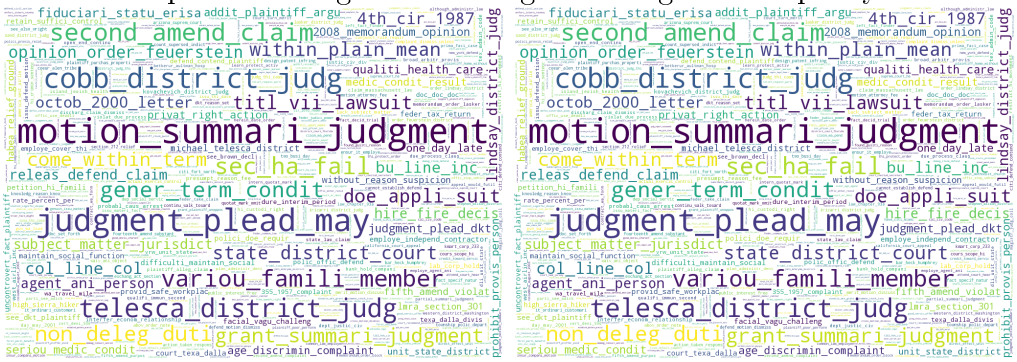


Figure 3: Most predictive bigrams and trigrams for racial disparity in sentencing



Next we determined the direction of the relationship (increasing or decreasing harshness) by the univariate correlation of the variable with the outcome. The n-grams shown in tables 4 through 6 break out the most predictive features by the predictive direction.

Figure 4: Most predictive bigrams and trigrams for gender disparity in sentencing



Overall the lists of predictive phrases are mostly not illuminating or interpretable. The phrases that are considered "important" do not seem to have much relation to language that would be considered bias. The most predictive bigram for reducing harshness is “non mandatory”, suggesting that the judge is discussing non-mandatoriness of sentencing, which may be used to justify discretion or lower sentences. There are also phrases discussing family (various family members, remove child), indicating concern with spillovers from sentencing.

For racial disparity, court opinions that contain the words “arkansa western division” or “Mississippi” are positive. These are both historically slave and Jim Crow states and point to institutional or cultural factors. In both the bigrams and trigrams tables, court opinions with word “Mississippi” were positive for racial disparity, meaning harsher sentencing for blacks relative to non-blacks. This could be picking up on the racial issues and implicit biases that may still be prevalent in the South.

8 Discussion and Next Steps

This paper reports promising preliminary results toward using the text of judicial decisions to measure judge preferences over policies. In particular, the judge opinion corpus may hold valuable information about a judges' bias and help predict disparity in sentencing. In future we see various ways we can improve our analysis.

First, we could try additional models. One attractive model in this context

Table 4: Phrases Related to Harshness

Increasing Harshness		Decreasing Harshness	
new_imag	motion_summari_judgment	non_mandatori	second_amend_claim
march_2004	judgment_plead_may	cir_2010	4th_cir_1987
retain_control	cobb_district_judg	decemb_2010	bu_line_inc
unit_state	telesca_district_judg	offic_stephen	octob_2000_letter
year_continu	sec_ha_fail	object_rais	lindsay_district_judg
total_absenc	variou_famili_member	remov_child	judgment_plead_dkt
commiss_texa	gener_term_condit	137_defend	one_day_late
test_test	non_deleg_duti	texa_dalla	age_discrimin_complai
record_owner	grant_summari_judgment	restrict_second	prohibit_provis_person
summari_judgment	state_district_court	walk_toward	without_reason_suspic
ann_arbor	titl_vii_lawsuit	prohibit_lawyer	feder_tax_return
standard_announc	within_plain_mean	adelphia_commun	court_texa_dalla
thi_case	doe_appli_suit	comptrol_gener	petition_hi_famili
app_635	come_within_term	march_2003	see_dkt_plaintiff
januari_1995	opinion_order_feuerstein	lindsay_district	uncontrovert_fact_plai
posit_show	col_line_col	deni_discharg	high_sierra_hiker
vagu_amorph	subject_matter_jurisdict	separ_unit	texa_dalla_divis
thi_court	agent_ani_person	act_intermediari	wa_travel_mile
rule_abov	releas_defend_claim	gap_exist	day_may_2001
shall_deduct	hire_fire_decis	water_manag	bank_hold_compani

Table 5: Phrases Related to Racial Disparity in Sentencing

Increasing Racial Disparity		Decreasing Racial Disparity	
ga_compani	natur_ga_compani	unit_state	motion_summari_judgm
section_new	market_channel_trade	summari_judgment	plaintiff_claim_relief
speed_limit	dure_interim_period	unsupport_probabl	unit_state_district
mississippi_defend	rule_motion_direct	march_2004	product_defend_argu
creat_substanti	arkansa_western_divis	motion_dismiss	without_leav_amend
singl_limit	keep_proper_lookout	district_court	properti_liberti_interest
mississippi_corpor	361_113_122	random_assign	unsupport_probabl_caus
mere_continu	tradit_maritim_activ	without_leav	detent_violat_hi
number_element	pose_secur_risk	opportun_cure	prove_set_fact
knowledg_reli	polici_wa_cancel	vehicl_park	state_district_court
safeti_hazard	demonstr_like_succeed	thi_court	without_prejudic_withou
decis_decemb	hire_fire_decis	thi_case	must_plead_factual
sec_research	preempt_feder_law	resciss_claim	wa_unsupport_probabl
fire_decis	procedur_state_summari	exhibit_list	outsid_unit_state
121_1511	wa_cover_insur	constitut_transact	worker_compens_benefit
review_submit	plu_function_form	would_adopt	client_prospect_client
surgeri_hi	review_prrb_decis	punish_student	jurisdict_count_iii
oklahoma_suprem	district_court_mississippi	feder_arbitr	claim_fals_arrest
2010_move	gener_charact_activ	161_supp	decemb_2004_court
tradit_maritim	charact_activ_give	contrari_public	file_june_2008

Table 6: Phrases Related to Gender Disparity in Sentencing

Increasing Gender Disparity		Decreasing Gender Disparity	
june_2001	alj_determin_claimant	deni_intervenor	182_fed_appx
mbna_bank	disput_resolut_provis	er_state	motion_summari_judgr
aircraft_oper	larsen_civil_right	unit_state	sentenc_thi_section
depart_taxat	bodili_injuri_coverag	summari_judgment	pay_minimum_wage
stop_limit	work_daili_basi	point_posit	offic_reason_suspicion
regularli_schedul	fair_busi_practic	busi_enterpris	dismiss_qui_tam
limit_coverag	marin_fisheri_serv	notic_sale	gener_maritim_law
feder_permit	658_4th_cir	2003_iep	unit_state_district
use_defens	profit_share_plan	doc_124	malpractic_claim_defer
mental_capac	complaint_fail_suffici	prepar_time	ei_arbitrari_caprici
test_period	jurisdict_defend_alleg	constitut_occurr	585_106_1348
use_provis	strong_infer_deliber	proof_applic	fraud_neglig_misrepres
state_oregon	prairi_band_potawatomi	close_end	set_asid_program
undu_influenc	decis_plaintiff_disabl	district_court	fraud_breach_fiduciari
mccarran_ferguson	decis_deni_coverag	actual_expend	limit_mean_plu
defend_constitut	hi_cell_door	worker_paid	hi_motion_attorney
resolut_provis	violat_ani_requir	hear_initi	see_fed_evid
jersey_superior	prudenti_in_america	pleasant_valley	llc_wa_form
distribut_busi	hog_mkt_inc	statement_statement	must_plead_enough
trust_deed	februari_2011_plaintiff	thi_case	accept_respons_action

could be partial least squares. We could also experiment with deep neural networks.

Second, we could experiment with other features – text, non-text, and their interactions. Text features could be further broken down into different parts of speech, proportion of negative and positive words, or deep learning methods such as Doc2Vec. The set of predictive features might be different for Republican and Democrat judges. Deep neural networks might do a good job in recovering the predictive power of these interactions.

The broader goal is to use this metric for studying legal and policy outputs. One could form the harshness or disparity predictions on other corpora of speeches or writings, not just judges but also attorney generals. One might expect that that the writings of Jeff Sessions (considered "tough on crime") would "predict" a higher harshness score than Eric Holder (who prosecutors more lee-way in charging criminals).

References

[ea15] Lim et. al. The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing Across Electoral systems. *American Economic Journal*, 7(4), 2015. 4.1

[Mus01] David Mustard. Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the US Federal Courts. *Journal of Law and Economics*, XLIV, 2001. 2

[NYT02] Wide Sentencing Disparity Found Among US Judges. *New York Times*, March 5, 2002. 2