Legal and Political Stance Detection of SCOTUS Language

Noah Bergam
Columbia University
New York, NY
njb2154@columbia.edu

Emily Allaway
Columbia University
New York, NY
eallaway@cs.columbia.edu

Kathleen McKeown
Columbia University
New York, NY
kathy@cs.columbia.edu

Abstract

We analyze publicly available US Supreme Court documents using automated stance detection. In the first phase of our work, we investigate the extent to which the Court's public-facing language is political. We propose and calculate two distinct ideology metrics of SCOTUS justices using oral argument transcripts. We then compare these language-based metrics to existing social scientific measures of the ideology of the Supreme Court and the public. Through this cross-disciplinary analysis, we find that justices who are more responsive to public opinion tend to express their ideology during oral arguments. This observation provides a new kind of evidence in favor of the attitudinal change hypothesis of Supreme Court justice behavior. As a natural extension of this political stance detection, we propose the more specialized task of *legal* stance detection with our new dataset `SC-stance`, which matches written opinions to legal questions. We find competitive performance on this dataset using language adapters trained on legal documents.

1 Introduction

The relationship between the Supreme Court of the United States (SCOTUS) and American public opinion is complicated. Some scholars debate normative questions as to whether the Court's power of judicial review ought to obey democratic principles[1] (Bassok and Dotan, 2013). Others investigate how SCOTUS behaves in relation to the public and why (Katz et al., 2017) . Prior work in the field of American political science has consistently demonstrated an association between the partisan ideology of the Court, as expressed through its decisions, and that of the public, as recorded through poll data (Casillas et al., 2011; Mishler and Sheehan, 1996, e.g.,). However, more recent work, particularly in light of the 2022 Dobbs v. Jackson decision, suggests a departure from this general pattern (Jessee et al., 2022). This change in institutional behavior has profound social significance which calls for academic attention. This paper heeds that call by providing a new analytical perspective on SCOTUS' democratic tendencies.

Despite extensive research confirming SCOTUS' general responsiveness to public opinion, the underlying reasoning for this relationship is disputed. One hypothesis centers on *strategic behavior*: it posits that the Court consciously acts in accordance with the public will in order to protect its Constitutionally fragile claim to the power of judicial review (Hammond et al., 2005). Alternatively, the *attitudinal change hypothesis* contends that broader socio-political forces such as news media present confounding factors that influence both the justices and the public (Norpoth et al., 1994).

In this paper, we gain new insight into these hypotheses by applying automated stance detection to a newly assembled corpus of Supreme Court written opinions and oral arguments. Stance detection (i.e., automatically identifying the position of an author towards a given target statement) allows us to evaluate the implications of a justice's language. We use stance detection and related techniques to build two different textual indicators of ideology which we call *issue-specific stance* (ISS) and *holistic political stance* (HPS) respectively. We compare these indicators to existing social scientific metrics related to general public opinion (i.e. the Stimson Policy Mood; Stimson, 2018), Supreme Court justice ideology (i.e. the Martin-Quinn score; Martin and Quinn, 2002), and Supreme Court case salience (i.e. the Clark case salience; Clark et al., 2015).

In addition, we build a supervised stance detection dataset, `SC-stance`, over a subset of Supreme Court written opinions. Our dataset matches the text of the written opinion to a corresponding le-

[1] Famously dubbed the "counter-majoritarian difficulty" by political scientist Alexander Bickel in 1962, this problem has been said to lie at the heart of American Constitutional scholarship (Friedman, 1998)

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 265 - 275

$D_1 =$	*Once the Court starts looking to the currents of public opinion regarding a particular judgment, it enters a truly bottomless pit from which there is simply no extracting itself.* (Rehnquist, 1992)
$D_2 =$	*Will this institution survive the stench that this creates in the public perception that the Constitution and its reading are just political acts?* (Sotomayor, 2022)
$T =$	*The Supreme Court ought to make decisions with the public opinion in mind.*
	stance(D_1, T) **= con** **stance**(D_2, T) **= pro**

Table 1: A relevant, sophisticated example of stance detection.

gal question (i.e., the target) posed on a legal educational website[2]. We present baselines on this dataset using tf-idf features, two language models for the legal domain (Chalkidis et al., 2020; Zheng et al., 2021), and a new method which involves augmenting BERT (Devlin et al., 2018) with an adapter (Pfeiffer et al., 2020a) pre-trained for the legal domain. We find performance gains both with this new method and from masking named entities in the training data.

The main contributions of this work are as follows. **(1)** Using stance detection, we formulate two distinct ideology metrics (i.e. *holistic political stance* and *issue-specific stance*) for SCOTUS justices serving from 1955 to 2020. We find that justices who are responsive to public opinion tend to use language which correlates ideologically with their voting behavior. This provides new evidence in favor of the attitudinal change hypothesis. **(2)** We release a new dataset, `SC-stance`, which matches written opinion text to relevant legal questions. It is the first *legal stance detection* dataset as far as the authors are aware. **(3)** We set baselines on our new dataset and find two ways to potentially improve performance: using a law-specific language adapter, and removing named entities during training.

The repository of relevant code is publicly available through the following link: https://github.com/njbergam/scotus-public-stance.

2 Related Work

Supreme Court and Public Opinion There is extensive academic work analyzing the Supreme Court's relationship with public opinion. In some cases, facts about the Supreme Court are gauged using a public opinion-related proxy. For instance, Segal and Cover (1989) developed an ideology score of justices based on newspaper editorials written at the time of their appointment while Epstein and Segal (2000) and Clark et al. (2015) used front-page news articles in order to quantify the political salience of Supreme Court cases. Other projects take a more direct look at the correlation between SCOTUS decisions and public opinion metrics. Casillas et al. (2011) uses a two-step least-squares regression approach in order to trace the public's influence on Court voting patterns, while Kastellec et al. (2010) looks at the relationship between state-level public opinion polls and Senator votes for SCOTUS justice nominations.

A common thread in many prior studies is the focus on Court voting behavior or its reception in the public eye. In contrast, our work investigates how SCOTUS presents its politics through its *language*. This approach takes advantage of the fact that the corpus of official SCOTUS language is publicly available, relatively small, and well-structured.

Previous work in various fields demonstrates that there are concrete differences between the language used by people of different political ideologies. In psycholinguistics, Robinson et al. (2017) suggests that the language of liberals tends to emphasize mental concepts, while that of conservatives uses more references to the body. NLP research has further investigated this concept through political ideology detection on two datasets (Iyyer et al., 2014, e.g.,): Convote (i.e. Congressional dialogue labeled with the political affiliation of the speaker) (Thomas et al., 2006a), and the Ideological Books Corpus (i.e. sentences from political articles and books annotated for political cues) (Sim et al., 2013).

Legal Artifical Intelligence The legal domain presents a unique challenge for NLP due to the precision, structure, and everyday importance of legal language (Dale, 2019). Furthermore, legal language is interesting in terms of its intersection with political discourse[3], a much more well-studied

[2] `0yez.org`

[3] This intersection can be problematic. The Code of Conduct for US Judges states: "A Judge Should Refrain from Political Activity" (Courts, 2019) and presents restrictions on language, e.g. no public endorsement of political candidates.

genre in NLP. In this work, we investigate that very intersection by leveraging existing stance and political ideology detection datasets in the context of legal language.

There are two major types of legal AI models (Zhong et al., 2020): rule-based methods, which are mostly supported by legal AI practitioners in industry, and embedding-based methods, which seem to garner the most attention from researchers in academia. The latter body of work has recently focused on adapting pre-trained language models (e.g., BERT) to the legal domain, either through law-specific pre-training, fine-tuning, or a combination of both (Chalkidis et al., 2020; Zheng et al., 2021). Due to the general accessibility of many legal documents around the world, a wide variety of legal NLP datasets are now available, six of which were recently consolidated into the LexGLUE benchmark (Chalkidis et al., 2021). Our dataset, `SC-stance`, provides a test of legal understanding which is not currently captured by existing datasets. Rather than evaluating the relevance between legal statements or documents, `SC-stance` goes a step further and tests the relative stance.

Stance Detection The task of stance detection is to determine the stance (e.g., `Pro`, `Con`, or `Neutral`) of a text on a target (e.g., 'abortion') (Mohammad et al., 2016) (see Table 1 for an illustration). In many works on stance detection, the topic is a noun-phrase (e.g., 'legalization of abortion') and texts are relatively short, such as posts from debate forums (Abbott et al., 2016; Walker et al., 2012; Hasan and Ng, 2014, e.g.,), and comments on news articles (Krejzl et al., 2017; Allaway and McKeown, 2020). Stance detection on Twitter towards political targets is particularly popular (Sobhani et al., 2017; Li et al., 2021; Cignarella et al., 2020; Lai et al., 2020; Taulé et al., 2017). Despite this interest, there is a lack of labeled stance data in the legal domain. Our dataset `SC-stance` not only fills this gap, it also challenges stance detection systems with complex targets (i.e., full sentences) and long documents (i.e., thousands of words).

3 Evaluating Political Stance

3.1 Methods

In the first phase of our work, we track how Supreme Court justices express political leanings in their public-facing language. We focus on two particular corpora: the set of written opinions (1789-2020), and the set of oral argument transcripts (1955-2020). The former was obtained through a Kaggle database (Fiddler, 2020) which used the Harvard CaseLaw Project's[4] API to collect full text files of 33,490 Supreme Court written opinions.The latter was scraped from the Oyez Project (Urofsky, 2001), a multimedia archive of SCOTUS data. We collected over 3.8 million lines of dialogue

Metric	Dataset	Model	$F1$	Acc.
ISS	VAST	Baseline	58.2	-
		Ours	62.8	63.4
HPS	Convote	Baseline	-	70.2
		Ours	75.3	76.3

Table 2: Performance of the stance detection classifiers. The baseline for VAST is a BERT-based model (Allaway and McKeown, 2020) and for Convote it is an RNN (Iyyer et al., 2014).

3.1.1 Linguistic Ideology Metrics

Stance detection allows us to represent the political polarity of judicial language through our two new ideology metrics: issue-specific stance (ISS) and holistic political stance (HPS). Both measure a speaker's ideology along the classic liberal-conservative spectrum (Stimson, 2012). However, they arrive at their answers very differently. The ISS evaluates a speaker's stance relative to a set of representative topics, while the HPS seeks to classify the political affiliation of the speaker directly. Both metrics are built on top of transformer-based text classification algorithms. Although the ISS and HPS are calculated by statement, it is understood that each requires large representative samples of a speaker's statements in to provide some insight into their overall ideology.

Issue-specific stance (ISS) To obtain a speaker's ISS, we gauge a speaker's stance on various liberal and conservative political statements. We adapt these statements from the Pew Political Typology Quiz (Center, 2021), which uses a variety of questions to evaluate ideology on a continuous scale from "Progressive Left" to "Faith and Flag Conservative." Based on how much the given text agrees or disagrees with each of the liberal and conservative statements (which are paraphrased for simplicity), we construct a score which is meant to gauge ideology.

[4] https://case.law/

If a higher score indicates a conservative leaning (this is, of course, an arbitrary choice), then we can frame the ISS calculation for a specific text t as follows. Given a set of targets which align roughly with liberal ideals S_L, a conservative counterpart S_C, and a stance model which maps to some signed interval $[-1, 1]$ we calculate ISS as follows:

$$\text{ISS}_{S_L,S_C}(t) = \sum_{l \in S_C} s(l,t) - \sum_{c \in S_L} s(c,t)$$

We formulate the above stance model as giving a continuous output. In practice, this amounts to adding the softmax probability of the predicted class, signed according to the ideology of the statement.

Holistic political stance (HPS) This metric seeks to immediately classify whether a given piece of language expresses more conservative or liberal ideology overall. As such, the underlying detector is not trained to detect stance relative to a specific topic; rather, it is trained to predict the ideology of the speaker. This framework may help provide a a broader psychological perspective on the underlying ideology of someone's language. For instance, suppose Robinson et al. (2017) is correct that liberals and conservatives generally express metaphors differently. Then HPS may pick up on that implicit ideological cue if it noticed such a pattern in its training data. In contrast, ISS is, by design, better at picking up on explicit cues such as the affirmation of a liberal or conservative belief. Additionally, HPS is simple to calculate (i.e., it is the confidence output of a binary ideology classifier). Unlike ISS, it does not require the parameters of liberal and conservative targets. This inherent simplicity also makes the HPS algorithm run faster.

3.1.2 Calculating HPS and ISS

ISS and HPS rely on pre-trained stance and ideology classification models, respectively. This means they require different datasets for training. For the ISS metric, we train a model using the Varied Stance Topics (VAST) dataset (Allaway and McKeown, 2020), which covers a large range of mostly political topics with broad themes like climate change and immigration. For the HPS metric, we train a classifier using the Convote dataset (Thomas et al., 2006b), which maps statements spoken by Congressional representatives to their partisan affiliation. We formulate both of these as binary stance classification tasks for the sake of simplicity.

Our classifiers use RoBERTa (Liu et al., 2019) without fine-tuning. As shown in Table 2, we achieve higher accuracy than the existing baselines in the original papers for each datset

To obtain the ISS or HPS score of a justice in a particular time period, we first collect the set of statements which contain some sort of emotion, with the intuition that this would increase the likelihood that the statement contains an opinion (as opposed to boilerplate legal language). To do this, we collect statements which feature a word from the NRC Emotion Lexicon (Mohammad and Turney, 2013). Then, for each justice, we collect a representative sample of statements per year and take the average score over all of these statements, to get the HPS of that particular year. In our experiments in the next section, we took sample sizes on the order of 10^3 per year per judge due to the time constraints of processing the text.

3.1.3 Baseline Metrics

We compare our linguistic ideology metrics to three existing metrics in the quantitative political sciences. These will serve as important baselines since they help us contextualize and evaluate our own metrics. These metrics have been calculated in previous research and are available through online databases.

Martin-Quinn scores[5] are dynamic ideal-point estimations of justices' political ideologies (Martin and Quinn, 2002). This metric, calculated on a yearly basis, uses a latent variable model where a justice's voting behavior is the observed variable.

The Stimson Policy Mood [6] gauges the general political leanings of the public through longitudinal surveys, which ask questions on a variety of issues over repeated time points (Stimson, 2018).

The Clark case salience[7] metric uses front page newspaper articles in The New York Times, The Washington Post, and The L.A. Times to quantify how relevant different Supreme Court cases are in the public eye (Clark et al., 2015).

3.2 Results

The first round of our analysis centers on the relationship between our linguistic ideology metrics and existing measures of Supreme Court behavior.

[5] mqscores.lsa.umich.edu/
[6] stimson.web.unc.edu/data/
[7] dataverse.harvard.edu/dataset.xhtml...

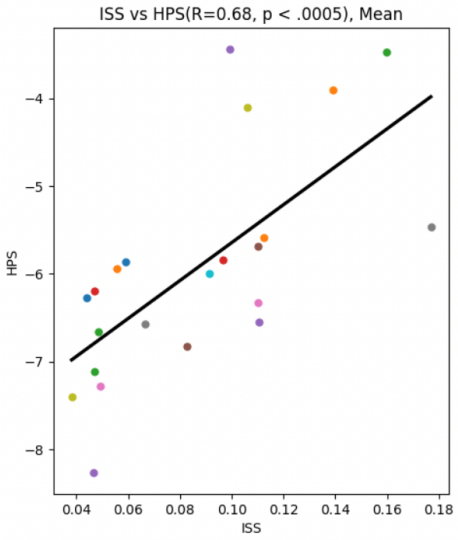

Figure 1: A strong correlation ($R = 0.68, p < 0.0005$) between the holistic and the issue-specific stance scores. Each data point represents a justice's mean score over their tenure (the significance drops to $p < 0.0001$ when we consider their median score).

ISS and HPS correlate. We first undertake a simple methodological audit and compare the issue-specific and holistic political stance scores across 23 justices who served from 1955 to 2020. We find that the two correlate quite strongly (Fig 1), despite the fact that the underlying stance detectors were formulated and trained in very different ways (§3.1). This suggests that the detectors are measuring the same signal and provides evidence that there is in fact a political signal in the dialogues of justices. This is not only important for our analysis, but it is also surprising in its own right given the officially apolitical stance of the Supreme Court (Courts, 2019).

Insight on the Attitudinal Change Hypothesis. Next, we looked at our metrics (ISS and HPS) in relation to the Martin-Quinn score. Importantly, we partition the justices based on their general responsiveness to public opinion. We measure this responsiveness by gauging the correlation between yearly Martin-Quinn scores (i.e., estimating justices' ideology) and the Stimson policy mood (i.e., estimating public opinion), by justice. We say that justices are "responsive" if this correlation is significant with $p < 0.05$.

We found that justices who are more responsive to the public opinion, compared to their counterparts, exhibit a much greater correlation between the ideology of their language, as measured by ISS and HPS and that of their voting decisions (Fig 2). This pattern is particularly noticeable with the HPS score. Additionally, this pattern intensifies when we looked purely at justices who have served past 1990.

This result offers new support for the *attitudinal change hypothesis*, which explains the correlation between Supreme Court decisions and public opinion by arguing that "the same social forces that shape the mass public also influence Supreme Court justices" (Casillas et al., 2011).

Our results support the *attitudinal change hypothesis* for two reasons. Firstly, note that a major underlying assumption of attitudinal change is that "individual attitudes are assumed to be the primary determinants of behavior" (Mishler and Sheehan, 1996). Thus, if justices are responsive to public opinion because of their attitudes, then these attitudes would affect both voting behavior and language. This is precisely what we observe when we find a correlation between Martin-Quinn scores and HPS for responsive justices.

Furthermore, the *strategic behavior hypothesis* does not have as much explanatory power for our results. HPS, by design, is sensitive to speech patterns that mirror those of Congresspeople. Considering the norms of the Court, it is more likely that such quasi-political behavior stems from latent, ideological influences rather than strategic behavior. If anything, strategic behavior would explain a correlation between ISS (i.e. explicit ideological expression) and MQ, which we did not observe.

Case salience and political language. We also consider political undertones of written opinions. We analyzed the relationship between the *magnitude* of the HPS of the written opinion text (a measure of its general political signal) and the Clark Case salience (i.e., public relevance) of the corresponding case. We found that the correlation was almost always slightly negative and only statistically significant for a handful of years (Fig 3).

This seemingly negative result actually parallels previous findings. In particular, Casillas et al. (2011) argue that public opinion may (counter-intuitively) hold less of an influence on salient cases as opposed to non-salient cases, since non-salient cases are simply more frequent. If the use of political language in a ruling can be seen as response to public opinion — which would seem to be the case under either of the leading hypotheses of Supreme Court behavior — then our result supports the the-

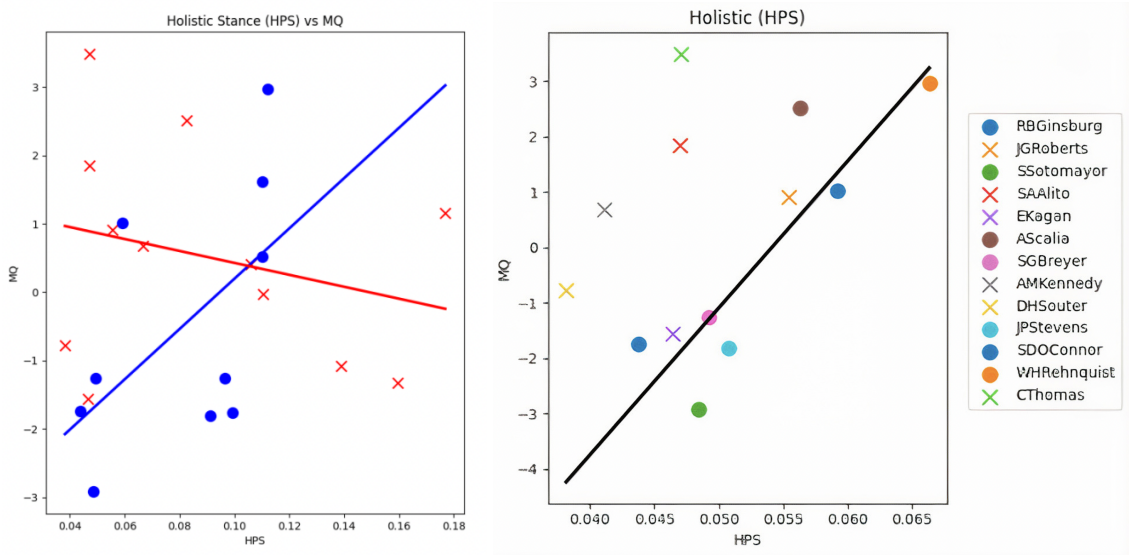

Figure 2: Mean holistic political stance versus mean Martin-Quinn score, by justice. In both figures, circles represent justices whose MQ scores correlate significantly with the Stimson policy mood over their tenure as justices. The left graph shows justices from 1955 to 2020, while the right shows and labels justices only after 1990. HPS was obtained using random sampling, with $n = 2000$ statements per year in the left graph and $n = 1000$ statements per year on the right.

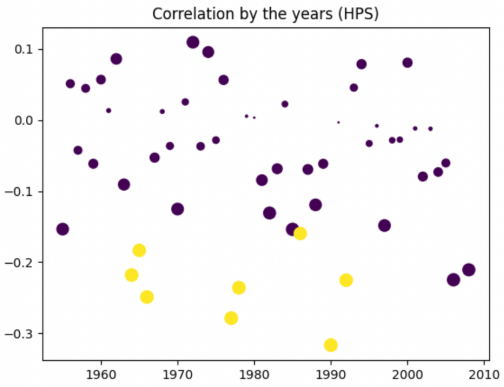

Figure 3: Correlation between confidence of the HPS score and the Clark case salience over all Supreme Court written opinions from 1955 to 2008. Yellow denotes a statistically significant correlation.

ory of an inverse relationship between salience and politicality.

4 SC-stance dataset

4.1 Methods

We describe the collection and characteristics of our new stance dataset, SC-stance, as well as the methods we apply to it.

Our dataset SC-stance was drawn from three sources: a dataset of full-text Supreme Court opinions through 2020 (Fiddler, 2020), the Washington University Supreme Court Database (Spaeth et al., 2014), and the Oyez website (Urofsky, 2001). We started by collecting written opinions which had non-neutral holdings, as encoded in the SC Database. We then automatically matched these opinion texts to the key legal question on the Oyez website to obtain text-target pairs. Since the questions on Oyez are always phrased such that an affirmative answer is in favor of the petitioner, we used the Winning Party label[8] from the Supreme Court Database, as well as the opinion type given in the Kaggle dataset (i.e. majority, concurring, dissenting, etc.) to infer the stance that a given written opinion takes towards the legal question (e.g. if the winning party was the respondent, and the opinion type was dissenting, then the opinion affirms the legal question).

The final dataset has 2708 labeled instances (1179 labeled pro, 930 labeled con). The average length of a target (i.e., the legal question) is 35 tokens and the average length of a text (i.e., the Supreme Court written opinion) is 5330 tokens. We show an example datapoint in Table 3.

In addition to providing a legal stance detection task, our dataset could provide an interesting passage retrieval task. Most other legal information retrieval datasets map documents to other docu-

[8] scdb.wustl.edu/documentation.php?var=partyWinning

Case: School District of Abington Township v. Schempp (ID 1962-148), Majority Opinion.
Target: Did the Pennsylvania law requiring public school students to participate in classroom religious exercises violate the religious freedom of students as protected by the First and Fourteenth Amendments?
Text: Once again we are called upon to consider the scope of the provision of the First Amendment to the United States Constitution which declares that "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof" [...] In light of the history of the First Amendment and of our cases interpreting and applying its requirements, we hold that the practices at issue and the laws requiring them are unconstitutional under the Establishment Clause, as applied to the States through the Fourteenth Amendment. [...]
Label: pro (text affirms the target)

Table 3: An example data point from SC-stance, in which we highlight the relevant portion of the text which confirms the stance.

ments (e.g., the German Dataset for Legal Information Retrieval (Wrzalik and Krechel, 2021)) or to static questions which are unchanged between documents (e.g., the Contract Understanding Atticus Dataset (Hendrycks et al., 2021)). The closest counterpart to our dataset, to the best of our knowledge, is the Belgian Statutory Article Retrieval Dataset, a French language dataset that maps legal questions written by laypeople to Belgian law articles (Louis et al., 2021).

4.1.1 Models for Stance Detection

In comparing models, we are most interested in which ones learn the most informative *features* from the text. The final layer is, in almost all cases, a single layer feed-forward network (Fig 4).

Legal Adapter Inspired by the concept that "legalese" could potentially be treated as a unique language, we use a *language adapter* to transfer a BERT-based stance detection model from its training data's domain to the SC-stance dataset. It is important to note that Supreme Court opinion language is relatively clear and concise compared to the more pure legalese of contracts or securities filings. While it may seem conceptually extreme to treat SCOTUS filings as a separate language, it is experimentally interesting as it sheds light on whether a dedicated adaptation for legal language allows for a more effective automated reading of legal stance.

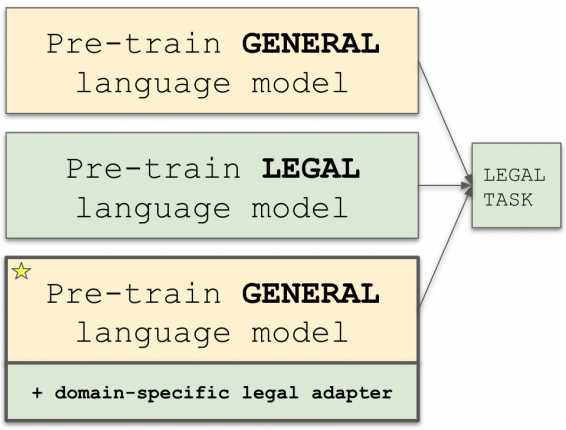

Figure 4: Three methods of tackling a legal NLP task using a large language model (the third being our new method which leverages language adapters). This paradigm generalizes to other domain-specific applications such as medicine or finance.

Adapters have been used to enable efficient multilingual transfer for language models. An adapter module is a set of weights (i.e., feed-forward layers) inserted into each attention block of a transformer and trained using masked language modeling (MLM). Adapters were originally designed as an alternative to fine-tuning (Houlsby et al., 2019) and have since become a popular method of cross-lingual domain transfer (Pfeiffer et al., 2020b; Vidoni et al., 2020, e.g.). One intuitive benefit of this approach over pre-training an entire language model is that only unlabeled data is needed to train the adapter and training is more parameter efficient, since the adapter has comparatively few parameters.

Baselines We compare our new method to a number of baselines, the simplest being the tf-idf vectorization of each of the target and document. On these simple features, we compare logistic regression (LR) and multilayer perceptron (MLP) as final layers; we find that the latter performs significantly better with $p < 0.02$[9], so we proceed to use MLP as the classification layer in our BERT-based models.

We experiment with BERT (Devlin et al., 2018), a popular transformer-based encoder pre-trained with masked language modeling and next sentence prediction. We also investigate two variants, which

[9]We use an approximate randomization test.

	Binary		3-class	
	Original	**w/ NER-mask**	**Original**	**w/ NER-mask**
Majority	39.6	-	20.4	-
tf-idf (LR)	41.4	43.2	26.5	29.6
tf-idf (MLP)	50.0	49.8	32.0	31.5
BERT	50.4	47.1	36.9	35.1
CaseLaw-BERT	47.6	49.2	38.3	40.3
Legal-BERT	52.8	53.0	**47.4**	41.7
Legal Adapter	**55.6**	**53.4**	41.4	**42.2**

Table 4: F-1 scores on the SCS-written dataset, using an 80-20 train-test split.

differ largely in terms of their training corpus. One is Legal-BERT (Chalkidis et al., 2020), which is pre-trained on an English legal corpus and uses a sub-word vocabulary built from scratch. The other is CaseLaw-BERT (Zheng et al., 2021), which is pre-trained on the Harvard Law case corpus.

4.1.2 Experimental Details

We evaluated our stance models on SC-stance in two settings: binary classification (i.e., labels {pro, con}) and 3-class classification (i.e., {pro, con, neutral}). Since SC-stance does not have any neutral labeled instances, following Allaway and McKeown (2020) we randomly pair opinions with unrelated questions to augment the dataset. For the adapter, we follow Pfeiffer et al. (2020a) and train a legal language adapter using MLM for $230k$ epochs with a learning rate of 10^{-4} and a batch size of 16. As unlabeled data we use over 8.8 million sentences from case law documents, made available through SigmaLaw (de Silva, 2019). In all experiments the SC-stance dataset is split 80/20 for training and testing. Importantly, we consider the case in which the training set has all named entities (with the notable exception of laws) masked during the training phase[10] and revealed during testing. This is referred to as the *NER-mask* setting in Table 4. For BERT and its variants, we append the legal question followed by a '[SEP]' token and the written opinion, and we truncate past the 512 token limit, with the understanding that most written opinions, despite their length, express their stance early on.

[10]We masked named entities using the Python spacy library's 'en_core_web_sm' model. The mask was the named entity type: for instance, "October 10" would become "[DATE] [DATE]".

4.2 Results

Overall, we found that the legal adapter is competitive with the leading legal language models, achieving the highest F-1 score (55.6) on the binary classification task[11]. In the 3-class setting, it was only outperformed by Legal-BERT.

We found that Legal-BERT consistently outperforms BERT and CaseLaw-BERT ($p < 0.09$ for the 3-class setting), which corroborates the experiments of Legal-BERT's creators (Chalkidis et al., 2020). We also found that, while the BERT-based features consistently outperformed the "classical" counterparts, the tf-idf model with an MLP classification layer had strong performance on the binary classification task.

We found mixed results with the NER mask setting, in that it led to both gradual increases (e.g. tf-idf with logistic regression, CaseLaw-BERT) as well as considerable drops in performance (e.g. legal adapter binary, Legal-BERT the 3-class setting). Intuitively, the NER mask should remove spurious signals for the classifier, since the relationship between the target and topic should almost never be related to the entities (i.e. proper nouns), but instead the relationships between entities.

We believe this hypothetical advantage is what led to certain score increases. However, the flipside is that there may be instability introduced when the model is presented with proper nouns in the test setting, after having had them removed during training. We noticed that BERT was more susceptible to this instability, which may be attributable to its less specialized vocabulary or understanding of legal grammar. These weaknesses of domain shift may increase the model's susceptibility to spurious signals.

[11]Due to the small size of the dataset, we were unable to mark these differences as statistically significant.

5 Conclusion

Using state-of-the-art NLP techniques, we gain new insight into a longstanding political science problem: the Supreme Court's relationship with public opinion. In our analysis of the language of Supreme Court justices, we leverage existing metrics of SCOTUS behavior as well as stance detection datasets regarding political ideology. Notably, we find a new source of evidence for the attitudinal change hypothesis of the Supreme Court, and we experiment with a competitive new model for legal language domain adaptation.

This research sheds light on how stance detection allows us to interrogate the implicit opinions of static documents. This is a powerful use case of NLP for the social sciences, in that it allows for a large-scale, critical analysis of large bodies of text. Of course, there is a long way to go in the field of stance detection, both generally and in specific linguistic domains such as the law. Our contribution of `SC-stance` feeds into this goal, by providing semantically rich targets and a mix of legal and lay language. We emphasize this latter feature, in that quality textual understanding – for human and AI alike – is marked by a thorough comprehension of both colloquial and technical language formulation.

Limitations

Our stance detection analysis of Supreme Court language is a proof-of-concept experiment with considerable potential for expansion. For instance, one could obtain a much richer understanding of Supreme Court ideology using a flavor of stance detection which analyzes targets relevant to issues of *jurisprudence* (e.g. judicial activism, originalism) rather than common politics.

There is also room for expansion in terms of our use and formulation of certain metrics. For instance, we chose not to investigate "public opinion" through text data, partly because the concept has no clear-cut representative corpus, and sampling from the web or the news could present selection biases. However, this problem could be resolved with a narrower view of public opinion such as, say, the news media. The inherent benefit to having a text-based metric of public opinion is that it is more easily comparable to text-based metrics of Supreme Court ideology. Furthermore, it may be enlightening to track the partisanship of justices' language using ideal point estimation (i.e., the words are the observed variable, the ideology is the hidden variable), rather than direct measurement of the justice stance year after year (Bafumi et al., 2005).

In terms of processing the `SC-stance` dataset, future work should look into how to work with the long written opinions using BERT-based methods which have a token limit. There is also clear potential to expand the `SC-stance` dataset. This could be done through strategic web-scraping of certiorari petitions, which often contain the relevant legal questions of (what eventually becomes) a Supreme Court case. If this challenge of locating the petitions, scraping the relevant text, and matching to the relevant case can be met, then the `SC-stance` dataset could in principle grow by orders of magnitude, which would make it an even more promising ground for experimentation.

Ethics Statement

Our investigation of the Supreme Court is an academic exploration of a political subject. By employing stance detection, we mean to uncover large-scale patterns in the text which may not be obvious to a single reader or scholar. This should not take away from the pursuit of engaging with text directly. After all, by transforming text into statistics, we lose many dimensions of its complexity in order to zero in on specific attributes. It is important to acknowledge this methodological complexity as quantitative social sciences research continues to engage with NLP-driven metadata.

Acknowledgements

The authors thank Hayley Cohen, Suresh Naidu, and the anonymous reviewers at NLLP for their comments. We thank Ariella Lang for organizing the Laidlaw Scholars Program which provided funding and support for this project. We also recognize the Columbia NLP Group for providing computing resources which made these experiments possible.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *LREC*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 8913–8931, Online. Association for Computational Linguistics.

Joseph Bafumi, Andrew Gelman, David K Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.

Or Bassok and Yoav Dotan. 2013. Solving the countermajoritarian difficulty? *International journal of constitutional law*, 11(1):13–33.

Christopher J Casillas, Peter K Enns, and Patrick C Wohlfarth. 2011. How public opinion constrains the us supreme court. *American Journal of Political Science*, 55(1):74–88.

Pew Research Center. 2021. Political Typology Quiz — pewresearch.org. https://www.pewresearch.org/politics/quiz/political-typology/.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *EVALITA*.

Tom S Clark, Jeffrey R Lax, and Douglas Rice. 2015. Measuring the political salience of supreme court cases. *Journal of Law and Courts*, 3(1):37–65.

United States Courts. 2019. Code of Conduct for United States Judges — uscourts.gov. https://www.uscourts.gov/judges-judgeships/code-conduct-united-states-judges#f. [Accessed 01-Oct-2022].

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

Nisansa de Silva. 2019. SigmaLaw - Large Legal Text Corpus and Word Embeddings — osf.io. https://osf.io/qvg8s/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lee Epstein and Jeffrey A Segal. 2000. Measuring issue salience. *American Journal of Political Science*, pages 66–83.

Garrett Fiddler. 2020. Scotus opinions. Full text and metadata of all opinions written by SCOTUS justices through 2020, https://www.kaggle.com/datasets/gqfiddler/scotus-opinions.

Barry Friedman. 1998. The history of the countermajoritarian difficulty, part one: The road to judicial supremacy. *NYUL Rev.*, 73:333.

Thomas H Hammond, Chris W Bonneau, and Reginald S Sheehan. 2005. *Strategic behavior and policy choice on the US Supreme Court*. Stanford University Press.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122.

Stephen Jessee, Neil Malhotra, and Maya Sen. 2022. A decade-long longitudinal survey shows that the supreme court is now much more conservative than the public. *Proceedings of the National Academy of Sciences*, 119(24):e2120284119.

Jonathan P Kastellec, Jeffrey R Lax, and Justin H Phillips. 2010. Public opinion and senate confirmation of supreme court nominees. *The Journal of Politics*, 72(3):767–784.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Peter Krejzl, Barbora Hourová, and Josef Steinberger. 2017. Stance detection in online discussions. *ArXiv*, abs/1701.00504.

Mirko Lai, Alessandra Teresa Cignarella, D. I. H. Farías, Cristina Bosco, V. Patti, and P. Rosso. 2020. Multilingual stance detection in social media political debates. *Comput. Speech Lang.*, 63:101075.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *FINDINGS*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Antoine Louis, Gerasimos Spanakis, and Gijs Van Dijck. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv:2108.11792*.

Andrew D Martin and Kevin M Quinn. 2002. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political analysis*, 10(2):134–153.

William Mishler and Reginald S Sheehan. 1996. Public opinion, the attitudinal model, and supreme court decision making: A micro-analytic perspective. *The Journal of Politics*, 58(1):169–200.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval@NAACL-HLT*.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Helmut Norpoth, Jeffrey A Segal, William Mishler, and Reginald S Sheehan. 1994. Popular influence on supreme court decisions. *American Political Science Review*, 88(3):711–724.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Michael D Robinson, Ryan L Boyd, Adam K Fetterman, and Michelle R Persich. 2017. The mind versus the body in political (and nonpolitical) discourse: Linguistic evidence for an ideological signature in us politics. *Journal of Language and Social Psychology*, 36(4):438–461.

Jeffrey A Segal and Albert D Cover. 1989. Ideological values and the votes of us supreme court justices. *American Political Science Review*, 83(2):557–565.

Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 91–101.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Harold Spaeth, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. 2014. Supreme court database code book. *URL: http://scdb. wustl. edu*.

James A Stimson. 2012. On the meaning & measurement of mood. *Daedalus*, 141(4):23–34.

James A Stimson. 2018. *Public opinion in America: Moods, cycles, and swings*. Routledge.

Mariona Taulé, Maria Antònia Martí, Francisco M. Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence. In *IberEval@SEPLN*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006a. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Matt Thomas, Bo Pang, and Lillian Lee. 2006b. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

Melvin I Urofsky. 2001. The oyez project: Us supreme court multimedia database. *The Journal of American History*, 88(2):753.

Marko Vidoni, Ivan Vulić, and Goran Glavaš. 2020. Orthogonal language and task adapters in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2012.06460*.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*.

Marco Wrzalik and Dirk Krechel. 2021. Gerdalir: A german dataset for legal information retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.