Supplemental Material
Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies[*]

Douglas R. Rice
Department of Political Science
University of Massachusetts Amherst
drrice@umass.edu

Christopher Zorn
Department of Political Science
Pennsylvania State University
zorn@psu.edu

February 20, 2019

[*] All materials necessary to replicate the results reported herein are posted to the *Political Science Research and Methods* Dataverse.

1 Size of Seed Set

While small changes in the size of the seed set are unlikely to be problematic, a few dynamics caution against regular expansions. First, given the differencing of the positive and negative word vectors in the calculation of most similar terms, the seed set is already effectively 20 seed words. Second, the most useful venues for additional seeds – better identification among small corpora – is complicated by the relative rarity of terms in those corpora. That is, selecting appropriate seeds in contexts with fewer terms becomes more complicated given the very shortage of terms. Conversely, and third, in venues with large corpora the value of additional terms is mitigated by the improvements in estimation of the word vectors in large corpora, thus additional terms risks biasing the estimates – given the difficulty of identifying *uncontroversially* positive or negative terms – in exchange for minor potential improvements in identification of similar terms.

2 Number of Extracted Words

We assess the accuracy of our approach across the number of words pulled for the dictionary. We estimate word vector representations based on the full 75,000 document set of positive, negative, and unlabeled documents. After estimating the vectors, we vary the size of the extracted positive or negative dictionary in increments of 50 from 100 to 1000 terms. We first extract 100 positive terms and 100 negative terms, then compute polarity and calculate classification accuracy within the corpus. The results appear in Figure 1. Most evident is the striking lack of variation across dictionary size, with the standard deviation of the series standing at 0.7%. Moreover, classification accuracy across the entire series is universally above that achieved by standard dictionary-based approaches, ranging from a maximum of 80.5% (200 positive and 200 negative terms extracted) to a minimum of 78.3% (900 positive and 900 negative). In all, there is strong evidence that the choice of the number of words to extract is of little consequence.

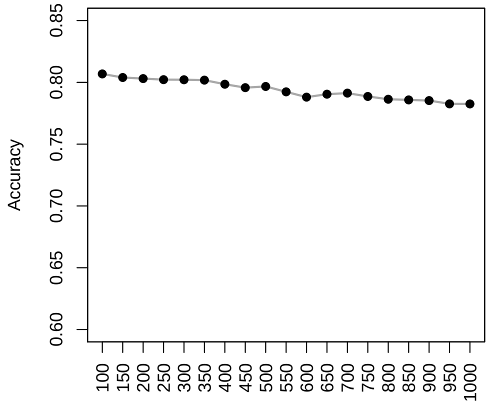

Figure 1: *Robustness To Differences in Size of Extracted Dictionary.* Plot of the accuracy (y-axis) of our polarity approach across variation in the size of the extracted dictionary (x-axis).

3 Distribution of Supreme Court Opinion Lengths

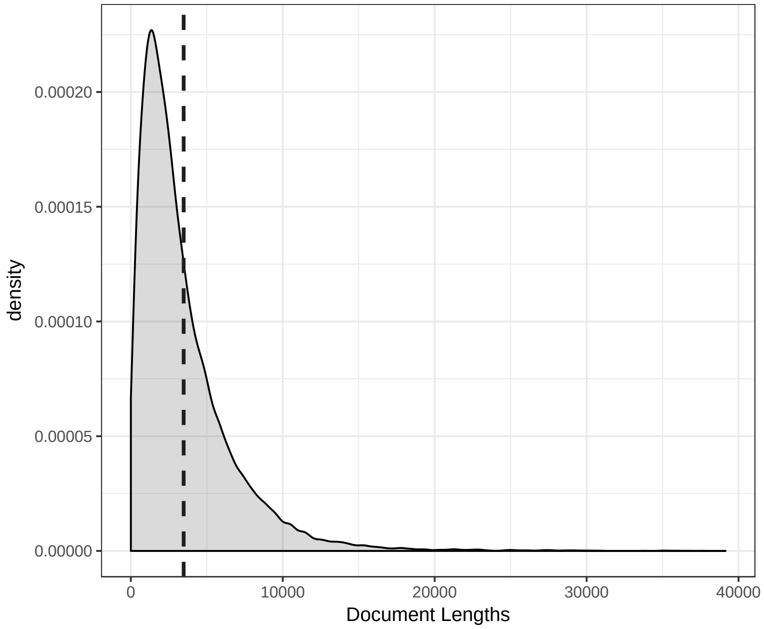

Figure 2: *Distribution of Document Lengths for Supreme Court Majority Opinions*.

4 Most Important Terms

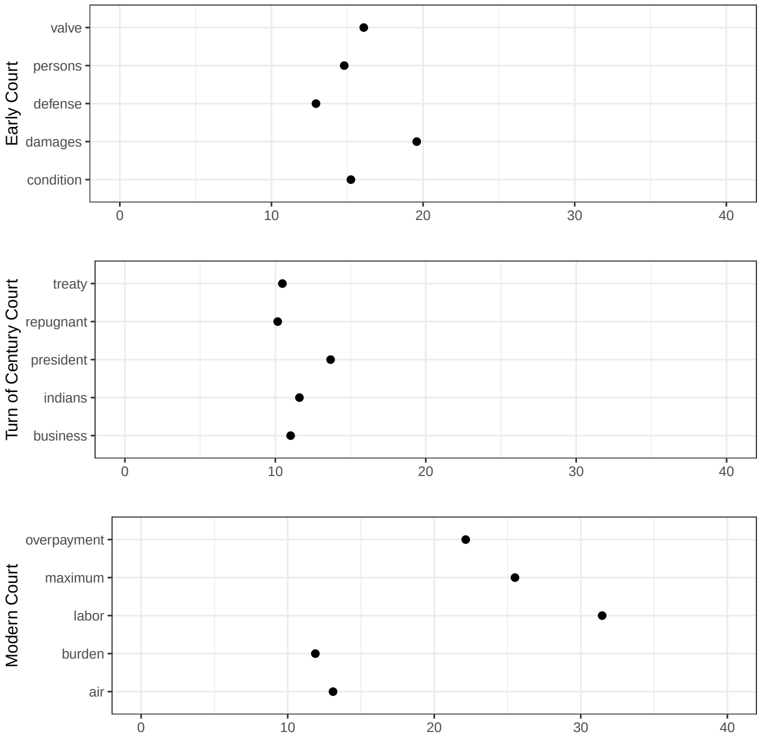

Figure 3: *Decrease in % Mean Squared Error.* Variable importance from random forest models of absolute difference between our polarity measure and LIWC-based polarity measure.

In this section, we seek to identify the most important terms in explaining the difference between our estimates and the LIWC estimates. We do so in order to better understand which terms our providing leverage across each era of the Court's history as we have defined it. For each era, we calculate the absolute value of the difference between our estimate and the LIWC estimate of polarity. Then, we predict the value using random forests regression (Breiman, 2001). We assess term (or variable) importance using the percent decrease in mean squared error. This is calculated by assessing model performance for the original model and permuted values over the dataset; where the performance is worse for permuted values one infers that the variable is more important. Notably, the terms are indicative of the ability of our approach to capture emotionally-valenced terms across eras that traditional, off-the-shelf dictionaries would miss. In the pre-1891 era, the terms include "persons", "damages", and "defense", around the turn of the century the terms include more topical items like "indians" and "treaty", and during the modern era the terms include "minimum" and "burden", again tokens that relate to particularly divisive topics.

5 Changes in Polarity Over Time

In Figure 4, we show the seven-year moving average of majority opinion polarity as estimated using our approach as compared to LIWC and AFINN estimates. Note first the trends of each line generally mirror one another; that is, our approach recovers sensible estimates of opinion sentiment. Yet the deviations in relative values are telling. The LIWC and AFINN estimates track more closely together, but with periods of stark disagreement; for instance, AFINN more closely aligns with our approach in the earliest periods before closely mapping to LIWC until approximately 1950, at which point LIWC begins a lengthy decline while AFINN identifies a relatively neutral Court until approximately 1990. Our approach, on the other hand, prior to 1925 identifies generally a marginally more divided Court, and a steep decline post-1925.

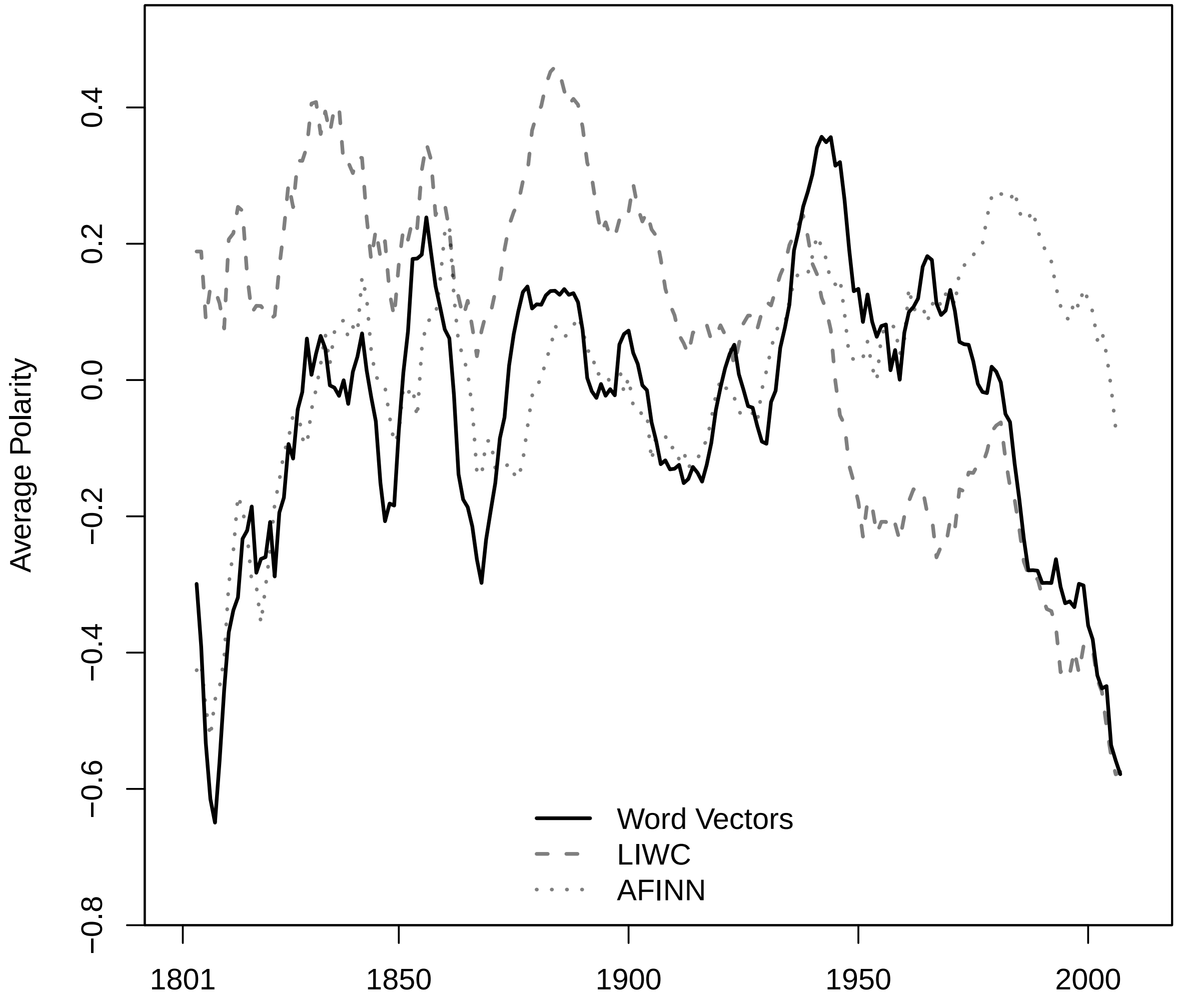

Figure 4: *Seven-year moving average of opinion polarity by year.* Plot of a seven-year moving average (symmetric) of average opinion polarity calculated using our approach (solid black line), the Linguistic Inquiry and Word Count dictionary (long dashed grey line), and the AFINN dictionary (short dashed gray line).

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543.