Political Science Research and Methods (2019), page 1 of 16
doi:10.1017/psrm.2019.10

ORIGINAL ARTICLE

Corpus-based dictionaries for sentiment analysis of specialized vocabularies

Douglas R. Rice[1]* and Christopher Zorn[2]

[1]Department of Political Science, University of Massachusetts, Amherst, Massachusetts, United States and [2]Department of Political Science, Pennsylvania State University, University Park, Pennsylvania United States
*Corresponding author. Email: drrice@umass.edu

(Received 12 October 2017; revised 23 April 2018; accepted 18 June 2018)

Abstract

Contemporary dictionary-based approaches to sentiment analysis exhibit serious validity problems when applied to specialized vocabularies, but human-coded dictionaries for such applications are often labor-intensive and inefficient to develop. We demonstrate the validity of "minimally-supervised" approaches for the creation of a sentiment dictionary from a corpus of text drawn from a specialized vocabulary. We demonstrate the validity of this approach in estimating sentiment from texts in a large-scale benchmarking dataset recently introduced in computational linguistics, and demonstrate the improvements in accuracy of our approach over well-known standard (nonspecialized) sentiment dictionaries. Finally, we show the usefulness of our approach in an application to the specialized language used in US federal appellate court decisions.

Keywords: Text and content analysis

Introduction

In the field of machine learning, an area of rapid recent growth is *sentiment analysis*, the "computational study of opinions, sentiments, and emotions expressed in text" (Liu, 2010). Broadly speaking, sentiment analysis extracts subjective content from the written word. At the most basic level, this might reflect the emotional valence of the language (positive or negative) but it can also entail more complex information content such as emotional states (anger, joy, disappointment). Tools for sentiment analysis allow for the measurement of the valenced content of individual words and phrases, sentences and paragraphs, or entire documents.

A number of approaches to estimating sentiment in text are available, each with benefits and potential risks. These methods fall into two broad classes. *Machine learning* approaches (e.g., Pang and Lee, 2004; Maas et al., 2011; Tang et al., 2014, 2016) rely on classifying a subset of texts (usually documents) on their sentiment, and then using their linguistic content to train a classifier; that classifier is subsequently used to score the remaining cases. In contexts where training data are available, machine learning-based approaches offer an efficient and accurate method for the classification of sentiment. These methods are less useful, however, in contexts without training data. These include many of the potential applications in the social sciences, where sentiment benchmarks are either entirely nonexistent, inappropriate, or difficult to obtain. In the latter instance, acquisition of training data typically requires the subjective human-coding of a substantial number of texts, an enterprise often fraught with unreliability. Failing that, the analyst

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

may only rely on previously-coded proxies believed to be reflective of sentiment. In either case, when no accurate training data are available, the application of supervised learning approaches introduces inefficiency and potential bias. Work by Tang et al. (2014, 2016) is illustrative; the authors build on recent research on word embeddings to learn not only semantic relations but also sentiment relations. However, the sentiment portion of their work provides a barrier to entry for many political science applications, as sentiment relations are learned by incorporating emoticons as learning outcomes. Such context-specific information for model training purposes is absent in most political science applications.

Alternatively, *dictionary-based* approaches begin with a predefined dictionary of positive and negative words, and then use word counts or other weighted measures of word incidence and frequency to score all the opinions in the data. With a completed dictionary, the cost for automated analysis of texts is extremely low (Quinn et al., 2010). As might be expected, though, the validity of such approaches turns critically on the quality and comprehensiveness with which the dictionary reflects the sentiment in the texts to which it is applied (Grimmer and Stewart, 2013). For general sentiment tasks, a number of pre-constructed dictionaries are publicly available, such as the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2001), and many have already found their way into published work (e.g., Black et al., 2011, 2016; Bryan and Ringsmuth, 2016). Pre-constructed dictionaries offer superlative ease of use. But while they have been applied across a variety of contexts, they are frequently context-dependent, potentially leading to serious errors in research (Grimmer and Stewart, 2013, 2). Conversely, constructing distinct dictionaries for each analysis is possible, but the costs of constructing a dictionary are often high (Gerner et al., 1994), and validating the dictionary can be difficult (Grimmer and Stewart, 2013).

Our goal is to develop an approach for building sentiment dictionaries for specialized vocabularies: bodies of language where 'canned' sentiment dictionaries are at best incomplete and at worst inaccurate representations of the emotional valence of the words used in a particular context. In doing so, we seek to maximize two criteria: the *generalizability* of the method (i.e., the breadth of contexts in which its application reliably yields a valid dictionary), and the *efficiency* of the method (in particular, the minimization of the extent of human-coding—and associated high costs—necessary to reliably create a valid dictionary). We propose and demonstrate the utility of a "minimally-supervised" approach to dictionary construction which relies on recent advances on measuring semantic similarity. Specifically, by identifying a small set of seed words correlated specifically with the dimension of interest in the domain and then—relying on word vector representations—computing semantically similar terms, one may extract a dictionary of terms which is both domain-appropriate and highly efficient. Across movie reviews and US Supreme Court opinions, we provide evidence of the efficacy of our approach and the associated improvements over extant methods.

Approaches to building sentiment dictionaries

The computational speed and efficiency of dictionary-based approaches to sentiment analysis, together with their intuitive appeal, make such approaches an attractive option for extracting emotion from text. At the same time, dictionary-based approaches have many limitations. Pre-constructed dictionaries for use with modern standard US English have the advantage of being exceptionally easy to use and extensively validated, making them strong contenders for applications where the emotional content of the language under study is expressed in conventional ways. At the same time, the validity of such dictionaries rests critically on such conventional usage of emotional words and phrases. Conversely, custom dictionaries developed for specific contexts are sensitive to variations in word usage, but come with a high cost of creation and limited future applicability.

What we term *specialized vocabularies* arise in situations when the standard emotional valences associated with particular words are no longer correct, either because words that typically convey emotional content do not do so in the context in question or *vice-versa*. For example, in colloquial English the word "love" almost always carries a positive valence (and its inclusion in pre-constructed sentiment dictionaries reflects this fact) while the word "bagel" does not. For professional and amateur tennis players, however, the two words might mean something very different; "love" means no points scored (a situation which has, if anything, a negative valence) and the word "bagel" refers specifically to the (negative) event of losing a set 6-0 (e.g., "putting up a bagel in the first set"). It is easy to see how the application of a standard sentiment dictionary to a body of text generated from a discussion of tennis could easily lead to inaccurate inferences about its content.

In such circumstances, an ideal approach is to develop a sentiment dictionary that reflects the emotional valence of the words as they are used in that context. Such dictionaries, however, are difficult, expensive, and time-consuming to construct, since they typically involve specifying every emotionally-valenced word in that context. The challenge, then, is to develop an approach for building sentiment dictionaries in the context of specialized vocabularies that is substantially more efficient and less costly than simple human coding.

Our approach to building specialized sentiment dictionaries leverages the structure of language and the corpus of text itself. That is, it constructs a dictionary from the words used in the very texts which are the subject of inquiry, and does so by relying on some universal facts about how words are used. Specifically, as our goal is to select a set of words related to a dimension of interest (here, sentiment), we seek to automatically identify broad sets of words related to that dimension. The intuition follows that of supervised learning, except that rather than coding the dimension across a set of training documents, we argue that by selecting a small set of terms ("seeds") we can grow a dictionary based only on identifying words occurring in similar contexts within the corpus.

Importantly, extensive prior work in natural language processing has focused on automatically identifying semantically similar words. In general, this research relies on the distributional hypothesis, and the idea that words used in similar contexts have similar meanings. Building from this central insight, researchers have recently sought to identify methods for understanding a word's *embedding* in a vector space; that is, these approaches seek to capture meaning that is lost in sparse, discrete representations of terms. Consider, for instance, the terms "king" and "queen". Standard approaches take the terms as discrete (i.e., 0 or 1). Instead, vector space models represent terms as distributions over word dimensions. Though none of the dimensions of the estimated vector are named, the "loading" of each term on the dimensions often captures substantively important relationships. For instance, "king" and "queen" might have a similar concentration on a dimension that seems to relate to the concept of *royalty* but deviate on a dimension that seems to relate to *man*. The resulting word vectors provide a wealth of linguistic information. By comparing or performing simple operations on the vectors, one frequently identifies semantically similar words or substantively interesting relationships (Mikolov et al., 2013). Though examples are myriad, a common version is to consider calculating the vector space of vec(`woman`) + vec(`king`) − vec(`man`), which results in a vector very similar to that of vec(`queen`).

Identifying the appropriate and relevant vector space, however, is difficult. Recent work by Mikolov et al. (2013) proposed doing so through shallow neural network models "useful for predicting the surrounding words in a sentence or a document" (2). The underlying idea is relatively straightforward; consider a target word surrounded by a certain number of context words. Next, predict the input word given the context words; Mikolov et al. (2013) perform this classification task using a single-layer neural network.[1] They find the estimated hidden layer from the neural network *by word* captures dimensions of semantic meaning.

[1]For simplicity, I focus here on the continuous bag of words [CBOW] variant. The skip-gram formulation flips the classification task, with the input word predicting context words.

A host of methods and extensions have been proposed through which to perform these calculations. One major innovation was the Global Vectors (GloVe) formulation. Here, the authors demonstrate the close connection between the word2vec method of Mikolov et al. (2013) and factorization of word co-occurrence matrices. In word2vec, the vectors are retrieved as a derivative of a classification task that relies on contextual word co-occurrences. GloVe instead is trained on *global* co-occurrence statistics. We utilize GloVe given initial work suggesting equivalent if not superior performance (e.g., Pennington et al., 2014) and its availability in `R`, while recognizing the ongoing debate as to when to employ word2vec or GloVe (see, e.g., Nematzadeh et al., 2017).

GloVe works as follows. Start by defining a word co-occurrence matrix X, where each entry X_{ij} indicates how often word i appears in the same context as word j.

Our goal is to construct word vectors for each word. To do so, for each word pair, GloVe defines a soft constraint such that

$$w_i^{\mathrm{T}} w_j + b_i + b_j = \log(X_{ij}), \tag{1}$$

where w_i is the word vector of the focus word, w_j is the word vector of the context word, and b_i and b_j are scalar bias terms. Then, to estimate the vectors, GloVe seeks to minimize a weighted least squares objective J:

$$J = \sum_{i=1}^{\mathrm{V}} \sum_{j=1}^{\mathrm{V}} f(X_{ij})(w_i^{\mathrm{T}} w_j + b_i + b_j - \log X_{ij})^2, \tag{2}$$

where V is the size of the vocabulary, and $f(X_{ij})$ is a weighting function which influences learning from common word pairs. The weighting function can take a variety of forms, but it is defined here (as in GloVe) as:

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{x_{\max}}\right)^{\alpha} & \text{if } X_{ij} < X\max. \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

In all, GloVe operates by directly examining *the ratio of word co-occurrence probabilities*. The model is trained using adaptive gradient descent (AdaGrad).[2]

As is clear throughout prior work, the estimated vector representations of words captures a great deal of meaning, and calculating the distance or angle between vector representations of words offers a method by which to recover semantically similar terms. We leverage this property to automatically construct dictionaries. Vector similarity is evaluated by cosine similarity. After estimating the vector representations for each token in our corpus, we identify positively-valenced tokens in the corpus by finding the token vector representations which are most similar to a vector constructed from the sum of a small set of 10 uncontroversially positive tokens *minus* the sum of a small set of 10 uncontroversially negative tokens.[3] These new words—semantically similar to the uncontroversially positive and negative seed words *within the domain*—are then extracted in order to construct the dictionary.[4]

The result is a pair of sentiment dictionaries—one comprised of positive tokens, one of negative tokens—that are derived from, and specific to, the corpus of text being analyzed. Yet beyond simple

[2]We utilize the `text2vec` (Selivanov, 2016) implementation of GloVe available in R.

[3]We identify the following seeds as uncontroversial across platforms. As positive terms, "superb", "brilliant", "great", "terrific", "wonderful", "splendid", "good", "fantastic", "excellent", and "enjoy." As negative terms, "bad", "awful", "badly", "dumb", "horrible", "wrong", "terrible", "poorly", and "incorrect."

[4]We additionally remove any terms which appear in either of the SMART stop words list, or that appear in the oppositely valenced dictionary of AFINN. We discuss changes in the size of the seed set in the online appendix.

counts of these terms, the process also provides a wealth of important information on the tokens; specifically, we know the distribution of usage across the corpus (term frequency—inverse document frequency [tf-idf])[5] and the similarity of the term's vector representation to the positive (negative) vector. Making use of this information, we weight token counts by tf-idf and then multiply the weighted counts by cosine similarity, yielding similarity weighted positive counts W^{p} and similarity weighted negative counts W^{n}. Letting T represent the set of tokens, polarity is then calculated as:[6]

$$\text{Polarity}_i = \frac{\sum_{t=1}^{\mathrm{T}} W_i^{\mathrm{p}} - \sum_{t=1}^{\mathrm{T}} W_i^{\mathrm{n}}}{\sum_{t=1}^{\mathrm{T}} W_i^{\mathrm{p}} + \sum_{t=1}^{\mathrm{T}} W_i^{\mathrm{n}}}$$

We think of this approach as "minimally supervised," in that it resembles in most respects unsupervised/data-driven approaches but requires at the outset a very small amount of human selection of the seed sets to serve as starting points for the learning algorithms. Our approach has a number of important benefits over supervised learning for many applications. First, the approach is domain appropriate while imposing minimal *a priori* structure on the corpus. The words are associated with the dimensions of interest only within the domain from which the texts were taken (addressing one primary concern of dictionary-based research) while also not being forced into potentially inappropriate classifications (a primary concern in supervised learning). Second, the approach is nearly costless compared to the alternatives. In the case of manually constructed dictionaries, selecting terms is exorbitantly expensive and validation is difficult. In terms of supervised learning approaches, there are extensive up-front costs in generating training data for classification and extensive validation. Third, and relatedly, the approach is much faster than the alternatives; the decrease in costs for generating dictionaries or training data is also associated with a massive decrease in the time necessary for implementation.

Validation: sentiment in movie reviews

We test our approach with the Large Movie Review Dataset.[7] These data consist of 100,000 movie reviews—25,000 positive, 25,000 negative, and 50,000 unlabeled—extracted from the Internet Movie Database (IMDB) archive. Positive or negative codes are derived from ratings provided by the reviewers. Prior research has utilized these or similar ratings extensively, primarily in the development of machine learning methods for the identification and measurement of sentiment in texts (e.g., Wang and Domeniconi, 2008; Maas et al., 2011). For our purposes, the assigned positive and negative ratings in the Movie Review Data provide a benchmark for assessing validity. An added benefit is derived from the fact that the sentiment of movie reviews is difficult to classify in comparison to other products (Turney, 2002; Dave et al., 2003). Thus, this application offers a difficult test for our approach to measuring sentiment, as well as the ability to precisely identify how accurate our approach is.

We begin by constructing word vectors from 75,000 documents: 12,500 positive, 12,500 negative, and the 50,000 unlabeled documents. The texts were stripped of punctuation, capitalization, and numbers. We drop the extremely frequent (the 20 most frequent tokens and any token appearing in more than 90% of documents) and extremely infrequent (appearing fewer than 90 times) tokens from the corpus. To create the co-occurrence matrix, we specify a context window of 50 tokens. To estimate the model, we use 300-dimensional vectors, setting $X_{\max} = 10$.[8]

[5]We employ tf-idf weighting in order to mitigate the influence of frequently used words.

[6]For purposes of comparison, we center and scale the polarity scores. This has a negligible impact on the accuracy of our approach but substantially improves the accuracy of both AFINN and LIWC.

[7]Available online at http://ai.stanford.edu/~amaas/data/sentiment/

[8]Recall that $X_{\max}$ is a critical value in the weighting function. Here, any word pair for which the co-occurrence count exceeds 10, the weight would be 1, whereas for all other word pairs the function returns a weight between 0 and 1.

We extract the top 500 positive and top 500 negative words by cosine similarity and calculate polarity according to the description above.[9] To calculate the accuracy of our approach, we employ a zero cutpoint, identifying all scores above zero as positive and all scores below zero as negative. With this cutpoint, we identify 11,845 reviews as negative and 12,745 as positive, with an overall classification accuracy of 80.2%.[10] For purposes of illustration and comparison, in Figure 1 we plot the estimated polarity of different dictionary-based approaches (x-axis) against the assigned ratings (y-axis), with a loess line demonstrating fit. Overall, our approach generally performs well, with the loess line shifting nearly perfectly at 0, as would be hoped. Moreover, it bears mentioning that inaccurately classified reviews disproportionately fall within the immediate region of the cutpoint, with reviews in this region frequently referencing the reviewers belief that the director or actors in the specific movie are typically good, but bad in the instant film.

As points of comparison, we also estimate polarity using two off-the-shelf dictionaries. The first is the LIWC software employed in prior work. Again defining zero as the cutpoint, LIWC correctly classifies just 69.4% of all movie reviews with scaling. Moreover, without scaling LIWC classifies more than two-thirds (67.8%) of movie reviews as positive. As our second comparison, we estimate polarity using the open-source AFINN-111 dictionary (Hansen et al., 2011; Nielsen, 2011), which provides pre-specified list of 2,476 words (1,598 negative and 878 positive) associated with scores between -5 (negative) and 5 (positive). Again, in Figure 1, we plot the associated ratings and classification. The figure provides stark evidence of the limitations of off-the-shelf dictionaries, as well as the difficulty of classifying movie reviews; overall, if we define "0" as the midpoint for the AFINN polarity measure, it classifies just 71.3% of reviews correctly. Loess lines plotted over each provide clear evidence of the improvement of our approach relative to standard dictionaries; the steep vertical ascent of the fitted line at 0 in the plot of our approach indicates the strong shift to classification of positive opinions as such, while neither LIWC nor AFINN approach similar shapes.

As a further check on the robustness of our approach, we also estimate polarity for a held-out set of 25,000 test documents, equally balanced between positive and negative reviews. While LIWC and AFINN are pre-defined dictionaries and thus accuracy should not shift substantially, our word vectors were "learned" from a separate set of documents. This therefore offers an additionally conservative test of the validity of our approach, as we take the dictionary estimated and extracted from one set of documents and apply it to another set of documents *of the same domain*.

As is clear from the bottom panels in Figure 1, the accuracy of each approach proves consistent across this new, held-out set of documents. Though expected in the case of AFINN and LIWC dictionaries, the ability of our approach to yield a dictionary applicable for held-out documents and at nearly identical levels of accuracy across sets offers important evidence of the validity and reliability of our approach. Words and estimates based on word similarity within the domain but not for the specific texts under study are, these results suggest, equally valid for estimation outside of the set with which they were estimated. In so doing, this offers strong evidence the recovered words and associated dimension are substantively valid representations of the concept of interest.

Robustness

In the following section, we compare our approach to the accuracy of a series of supervised learning alternatives, demonstrating yet further the benefits of building a dictionary through a small seed set of terms and identification of semantically similar word vectors. Before doing so, however, we assess the robustness of our approach across corpora size.

[9]Results are robust to variation in the number of the extracted words. See online appendix for details.
[10]The remaining 410 are classified as neutral as they feature no positive or negative words.

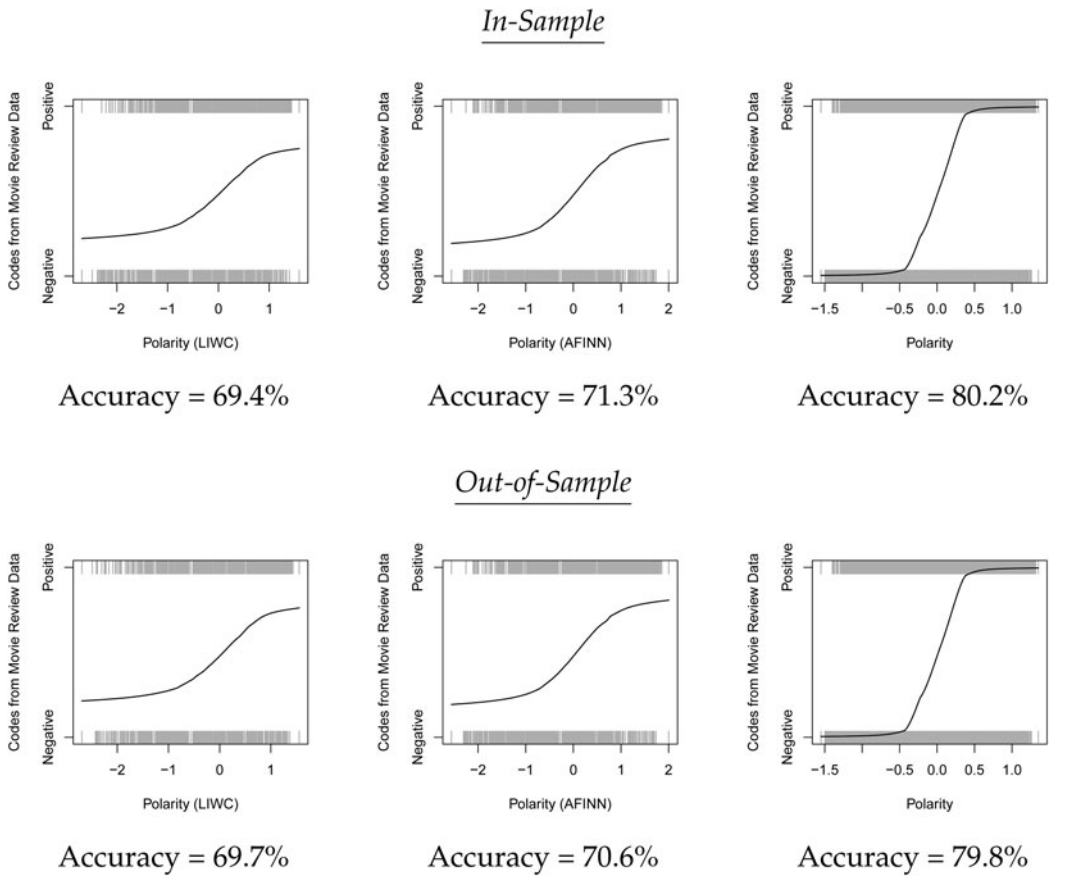


Figure 1. *Accuracy of Classifiers.* This compares three different approaches to measuring polarity against the assigned ("true") classifications. The left-hand panel compares estimates derived from the LIWC dictionary, the middle panel estimates from the AFINN dictionary, and the right panel estimates from our corpus-based dictionary The top row indicates cases within the sample used for training the word vector representations, whereas the bottom row indicates accuracy on held-out cases.

The most important dimension of corpus size for word embedding tasks is the total number of tokens, which is a function of the length of the document and the total number of documents. While the movie reviews corpus features an enormous number of documents, the length of those documents is modest. In the left panel of Figure 2, we plot the distribution of document lengths. Notably, the overwhelming majority of these documents feature fewer than 500 tokens. The average (mean) document is 184 tokens long, with a maximum value of 2,085 tokens.

Yet though the documents are generally short, our approach quickly achieves high levels of accuracy as the number of documents increases. To see this, we estimate a series of models across variations in the number of documents in the corpus. To do so, we use parameters consistent with those employed above.[11] Each iteration is a sample of the 75,000 document corpus, meaning each sample includes positive, negative, and unclassified opinions. Unclassified opinions are retained because they arguably introduce a harder challenge and more conservative assessment of our approach; though many of these certainly may be positive or negative, others are likely more neutral in character than those for which sentiment rating was provided. Accuracy is

[11]We shift the minimum number of word occurrences by a common ratio across iterations.

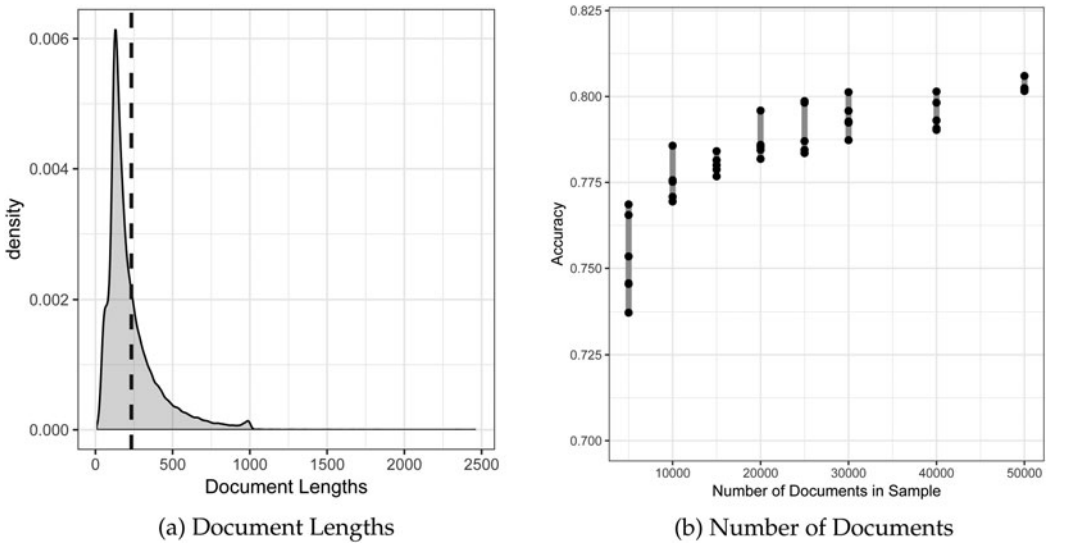


Figure 2. *Robustness To Differences in Corpus Size.* Plot of the accuracy (y-axis) of our polarity approach across variation in the corpus size (x-axis).

assessed within each sample, with the neutral reviews excluded in computing accuracy. At each sample size, we iterate through five samples to provide some evidence of estimation uncertainty.

The results are presented in Figure 2. A number of important dynamics jump out. First, even when the corpus is restricted to very small samples (5,000 documents), our approach outperforms the accuracy reported above for LIWC and AFINN. Whereas vector representations are regularly trained on large corpora of hundreds of thousands and millions of documents, even in a small-scale setting the identification of semantically similar terms offers an improvement on standard dictionary approaches utilized in social science research. Second, and as expected, as the size of the corpus grows so too does the accuracy of the classification. Moving from 5,000 to 15,000 documents yields large increases in accuracy and decreases in uncertainty, while each additional gain of 10,000 documents after 30,000 yields yet further marginal increases in accuracy. Such is to be expected; larger corpora provide greater information with which to identify accurate vector representations and likewise to derive appropriate dictionaries. Finally, as we discuss in our illustrative example, these results also have important implications for scholars interested in estimating sentiment from language over long periods of historical data. Because the models quickly exceed standard approaches to estimating sentiment, one can estimate separate models across smaller sets of documents in the study of sentiment over extended periods of time. In doing so, our approach addresses the well-known phenomenon of semantic drift which has vexed historically oriented text-as-data research.

Comparison to supervised learning

Before we demonstrate the utility of our approach in a particularly difficult research setting, it is imperative to note here that we do not argue that our approach is a universal substitute for supervised machine learning of sentiment. Such methods offer a useful tool to the classification of sentiment in texts *when clear benchmarks exist on which to train the classifiers*. But, in research areas where no natural benchmark is available for training a classifier, researchers are left with the unenviable task of developing coding protocols for often—in the case of the social sciences—lengthy texts with sophisticated or context-specific speech. Each of these components would

necessarily complicate the process of developing a reliably and validly coded set of texts of sufficient magnitude to develop and test supervised classifiers, and would likewise and relatedly carry high time and resource costs. Therefore, we believe it is important to emphasize the utility of an alternative method that approaches or exceeds supervised learning accuracy rates while being much less expensive and much faster to implement.

While our approach clearly achieves the latter two, the former—the accuracy of our approach relative to supervised learning implementations—deserves attention. To demonstrate, in Table 1 we compare our accuracy to seminal works in supervised learning for sentiment analysis of movie review data. In Pang et al. (2002), the authors employ a series of now-standard machine learning classifiers to the original movie reviews dataset, generally achieving accuracy rates approaching 80% across classifiers. In a development to that research, Pang and Lee (2004) introduced an approach for jointly modeling subjective terms and sentiment, yielding an increase in predictive accuracy of approximately 7%. In the most recent research, Maas et al. (2011) utilize vector representations of words to jointly model semantic content *and* movie review sentiment, providing minor improvements and raising overall accuracy to approximately 88%.

By comparison, our polarity approach achieves 80% accuracy, falling approximately in line with common, standard machine learning approaches. Moreover, and as documented above, the classification accuracy is consistent across the size of extracted dictionaries, and achievable in line with standard machine learning approaches once the corpus reaches approximately 15,000 documents. Though our approach falls short of two recently introduced methods, it does so *with no information on classification*. That is, while each of the supervised approaches explicitly utilizes the assigned classifications to identify features and mappings in order to optimize classification accuracy, our dictionary-based approach has no information on the outcome of interest. That such an approach yields estimates close to the best-performing machine learning classifiers—and indeed equals the success of many commonly employed classifiers—provides strong evidence of its utility to researchers. Having documented validity, we turn next to a unique domain.

Illustration: sentiment in US Supreme Court opinions

In the study of the US Supreme Court, a long trajectory of research has focused on the degree of consensus among the justices. A great host of questions animates this research, tracking from the influence of dissent on the Court's impact (Salamone, 2013), on public acceptance of the decision (Zink et al., 2009), and on legal development (Epstein et al., 2011; Rice, 2017). Yet further, the underlying question of how divided the Court is undergirds the long debates in judicial politics over the decline in the norm of consensus and the role of the chief justice in precipitating that decline (e.g., Danelski, 1960). Specifically, the norm of consensus refers to efforts on the part of justices to keep private whatever disagreements they might have, and thus to present to the public the image of a generally unified Court. In so doing, divided votes were much more rare in earlier periods not necessarily because the justices all agreed with one another but rather because a norm cautioned against airing disagreements in public. Throughout, then, a central challenge has been the measurement of comity on the Court; researchers have tended to rely primarily on the writing of concurring and dissenting opinions (e.g., Walker et al., 1988; Haynie, 1992; Caldeira and Zorn, 1998; Hendershot et al., 2013), but the existence of consensual norms—again, masking private disagreements from the public—make it likely that such indicators will significantly understate the true level of disagreement on the Court, and likewise be poor indicators of the effect of disagreement on many of the dynamics of interest to law and courts scholars.

One possibility, then, is to rely instead on the texts of the opinions. Opinion language is the central mechanism by which justices convey the substance of their rulings to the legal community and the public. Yet the opinions also contain language that—often strongly—conveys their

Table 1. Accuracy of Machine Learning Classifiers and Our Polarity Approach

Model	Mean	Min	Max
Pang et al. (2002)			
Naive Bayes	79.7	77.0	81.5
Maximum Entropy	79.7	77.4	81.0
Support Vector Machines	79.4	72.8	82.9
Pang and Lee (2004)			
Subjectivity SVM	87.15	–	–
Maas et al. (2011)			
Supervised Word Vectors	88.05	87.3	88.89
Our Approach			
Our Polarity Approach	80.2	–	–

emotions. Consider *Moran v. Burbine*[12] (1986), which dealt with Fifth and Sixth Amendment rights of the criminally accused. Writing for the majority, Justice Sandra Day O'Connor stated Burbine's argument would upset the Court's "carefully drawn approach in a manner that is both unnecessary for the protection of the Fifth Amendment privilege and injurious to legitimate law enforcement" while also finding the "respondent's understanding of the Sixth Amendment both practically and theoretically unsound." Dissenting, John Paul Stevens called the Court's conclusion and approach, "deeply disturbing," characterized the Court's "truncated analysis…(as) simply untenable," expressed concern that the "possible reach of the Court's opinion is stunning," and stated that the "Court's balancing approach is profoundly misguided." Responding, O'Connor's majority opinion stated that "JUSTICE STEVENS' apocalyptic suggestion that we have approved any and all forms of police misconduct is demonstrably incorrect." In footnotes, O'Connor went further, stating that the dissent's "lengthy exposition" featured an "entirely undefended suggestion" and "incorrectly reads our analysis." In footnote 4, O'Connor states "Among its other failings, the dissent declines to follow *Oregon v. Elstad*, a decision that categorically forecloses JUSTICE STEVENS' major premise …. Most importantly, the dissent's misreading of *Miranda* itself is breathtaking in its scope."

As is clear above, divisions on the Court regularly find their way into the written words of the justices. However, there is no readily-accessible approach for machine-coding the sentiment of judicial opinions. We instead utilize our approach. To undertake this analysis, we acquired the texts of *all* Supreme Court cases from 1792 through 2010 from justia.com, an online repository of legal documents. To get opinion-level data, we wrote a computer program which separated each case file into separate opinion files and extracted information on the type of opinion (majority, concurring, dissenting, special, and per curiam) and the author of the opinion. To maintain comparability across eras with vastly different separate opinion writing practices, we retain only majority opinions for these analyses. We then matched the opinions to the extensive case information available from the Supreme Court Database (Spaeth et al., 2012). Texts were cleaned according to standard text preprocessing steps, though note that terms were not stemmed.[13] The data thus constitute a comprehensive population of the majority opinions of Supreme Court justices, with nearly 26,000 unique opinions spanning more than 217 years of the Court's history.

As noted previously, a vexing problem for text-as-data classification across long periods of time—as here—is the issue of semantic change. One might reasonably worry that words with a negative valence in 2000 may not have the same negative valence, or may even be positively valenced, at some earlier period of history. Our approach offers a fast and flexible method of

[12]475 U.S. 412.

[13]Specifically, we removed capitalization, numbers, and punctuation.

addressing this semantic shift.[14] Specifically, because our approach to automatically deriving dictionaries from the corpus achieves high accuracy rates even in small corpora, one can examine sequentially different subsets of the corpora, thereby accounting for semantic shift over the long range. Here, we split the corpus into three subsets based on historical understandings of shifts in the Court's role: first, all opinions written before 1891 or the date of the Evarts Act, which fundamentally shifted the role of the Court; second, all opinions written between 1891 and 1925, or the date of the Judges Act and a common point at which researchers claim the Court's norm of consensual behavior begins to waver; and finally all cases after 1925. Though this partitioning of the corpus decreases the number of documents available for each model, recall that the chief concern for embedding models is the number of tokens. On this front, the length of Supreme Court opinions proves fruitful, as the *average* (mean) opinion length is 2,097 tokens, *longer than the very longest* movie review.[15]

We applied our approach across windows to estimate the aggregate sentiment of each opinion. The value of unique estimates across different windows is evident in the obtained dictionaries and changes in the terms identified as emotionally valenced. On this front, overlap between the estimated dictionaries is relatively modest. Going from the early Court to the turn of the century Court, only 48.8 percent of negative terms and 36.5 percent of positive terms are retained. Going from the turn of the century to the Modern Court, only 35.7 percent of negative terms and 22.1 percent of positive terms are retained. Thus, significantly different terms are captured across each window, to say nothing of the information available in the weighting across these windows.

Turning then to validation of our estimates, we approach the performance of our measure across two criteria: first, the degree to which they correlate with other variables they should theoretically be correlated with (convergent validity), and second, the degree to which they diverge from alternative measures in theoretically important ways (discriminant validity). We begin with convergent validity. Prior work regularly employed voting divisions as evidence of the social environment (e.g., Walker et al., 1988; Haynie, 1992). Though imperfect for reasons stated above—notably, the existence of consensual norms in earlier periods of the Court's history—the voting division measure offers a very coarse picture of the Court's level of disagreement. Specifically, in a seminal piece, Danelski (1960) argues the social environment of the Court is shaped by the degree to which social leaders emerge to minimize differences, relax tensions, and improve social relations in the small-group context.

To see this, in Figure 3 we plot the mean majority opinion polarity across different values of the number of dissenting votes; again, though not a perfect analog the public expression of disagreement should correlate with our measures of opinion polarity if those measures capture the Court's latent disagreement.

The results are illustrative of the value of our approach. The top panel provides mean opinion polarity calculated across the entire corpus, and reveals interesting similarities and differences in the estimates. Both our measure and AFINN indicate that unanimous opinions are generally neutral in tone. Where one justice dissents, the opinion polarity of the majority opinion is actually slightly positive, on average. From there, however, the values begin to tail off, eventually moving to negative. Contrast this with LIWC, which quickly moves to extreme negative values but with a slight increase atthree dissenting votes. The differences in approaches are most stark, however, among subsets of the Court's history. In early terms, all three measures show increases in majority opinion polarity as the number of dissenting votes goes up. Note, however, that LIWC remains at large positive values across the full range.[16] Contrast this with LIWC in the modern era, where

[14]Here we motivate the temporal splits in the data at theoretically-appropriate junctures in the Court's history. Scholars especially interested in semantic change might instead consider estimating the dictionaries across rolling windows.

[15]A density plot of opinion lengths appears in the appendix.

[16]Given changes in the size of the Court in the early era, there are few instances with four dissenting votes.

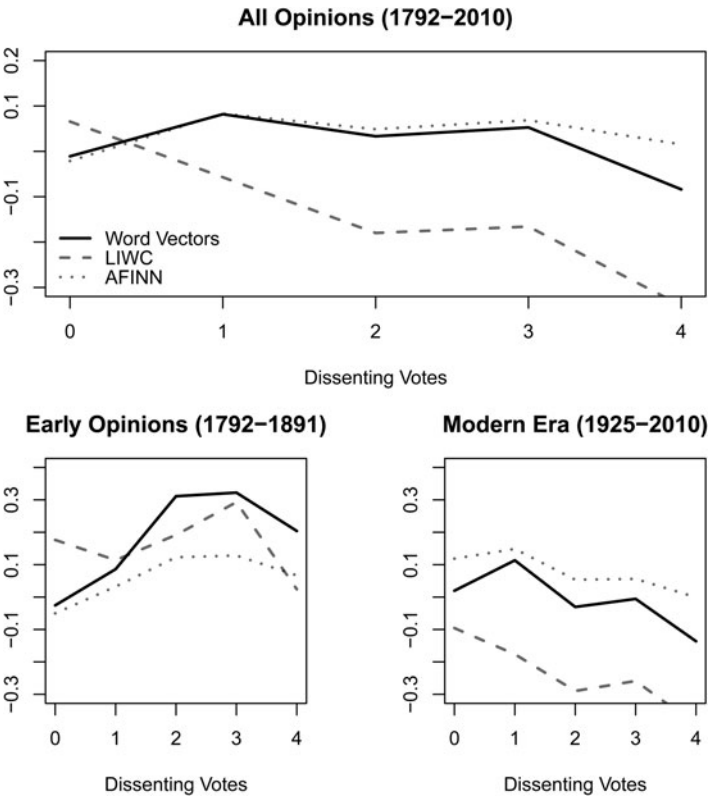


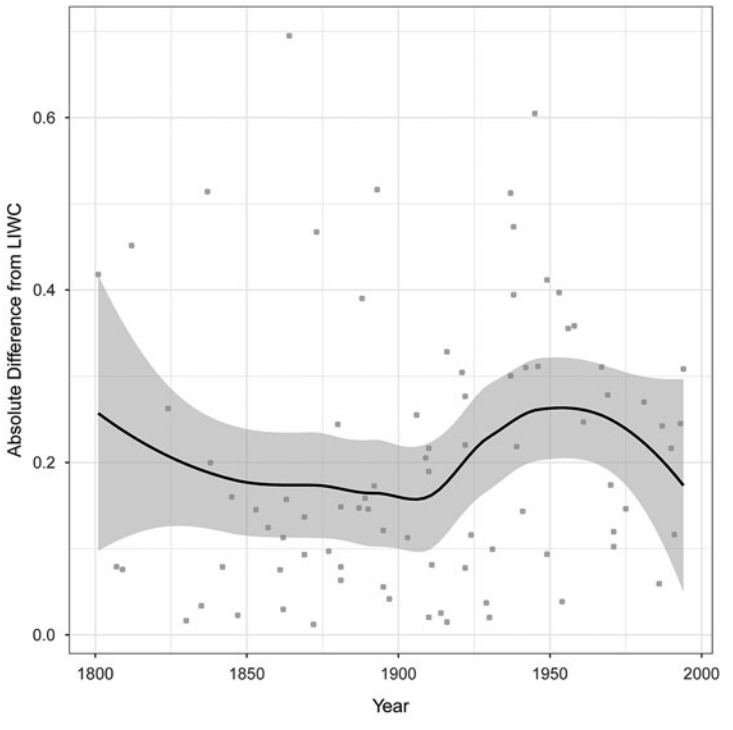
Figure 3. *Average polarity of majority opinions across different values of dissenting votes for different eras of the Court's history.* Plot of mean opinion polarity (y-axis) by number of dissenting votes (x-axis) calculated using our approach (solid black line), the Linguistic Inquiry and Word Count dictionary (long dashed gray line), and the AFINN dictionary (short dashed gray line).

it takes on large negative values across the full range of voting divisions. Our approach, though, again yields a generally neutral unanimous majority, followed by a slight increase and a subsequent steady drop. In the modern era—when standard dictionaries should work best—our approach yields estimates that are generally correlated but largely more sensible.

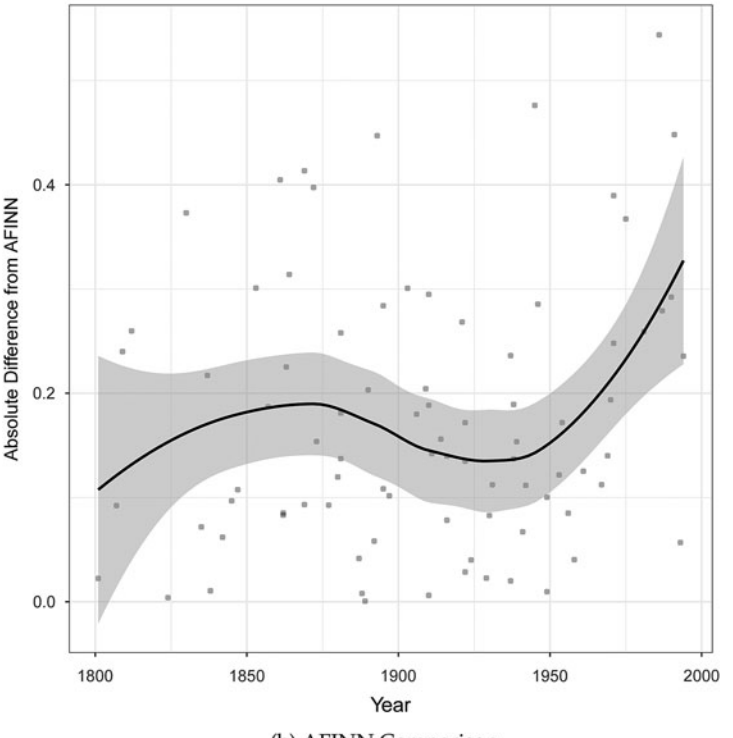
For discriminant validity, we turn our focus to changes in our understanding of individual justices. To do so, we calculate the average majority opinion polarity for each Supreme Court justice who authored more than 50 majority opinions, then compute the difference between our measure of polarity and measures derived from the AFINN and LIWC dictionaries. In Figure 4, we plot the absolute difference of our measure with that of LIWC (top panel) and AFINN (bottom panel) by the term in which the justice joined the Court. The local polynomial fit lines are instructive; in both plots, an inflection point occurs almost precisely at 1925, after which the divergence between with LIWC decreases while the divergence with AFINN increases. In contrast, prior to 1925 the divergence between our measure and LIWC is generally consistent, with a short decrease in early years of the Court, while our divergence with AFINN increases until the mid-19th century.

At the individual justice level, the value of our approach is also clear. In Figure 5, we plot the five justices with the largest positive (top five rows) and the five justices with the largest negative shifts (bottom five rows) in polarity across our dictionary and the LIWC (left panel) and AFINN (right panel) dictionaries. To contextualize the magnitude of the differences across measures, the standard deviation among justice-level mean polarity is 0.22 for our polarity measure, 0.23 for the LIWC-based measure, and 0.16 for the AFINN-based measure. Thus, the differences in estimated justice-level polarity is generally on the order of two standard deviations in the overall distribution of justice-level polarity.

The results tell a number of important stories about the history of the Court and the validity of our measurement strategy. Consider, first, the major positive switch for Justice Harold Burton



(a) LIWC Comparison



(b) AFINN Comparison

Figure 4. *Over Time Difference in Polarity Estimates* The above plots the absolute difference in justice-level polarity averages between common dictionary approaches and our estimate (y-axis) against the justice's term of arrival to the Court (x-axis). Black line represents local polynomial fit with associated 95% confidence interval in gray.

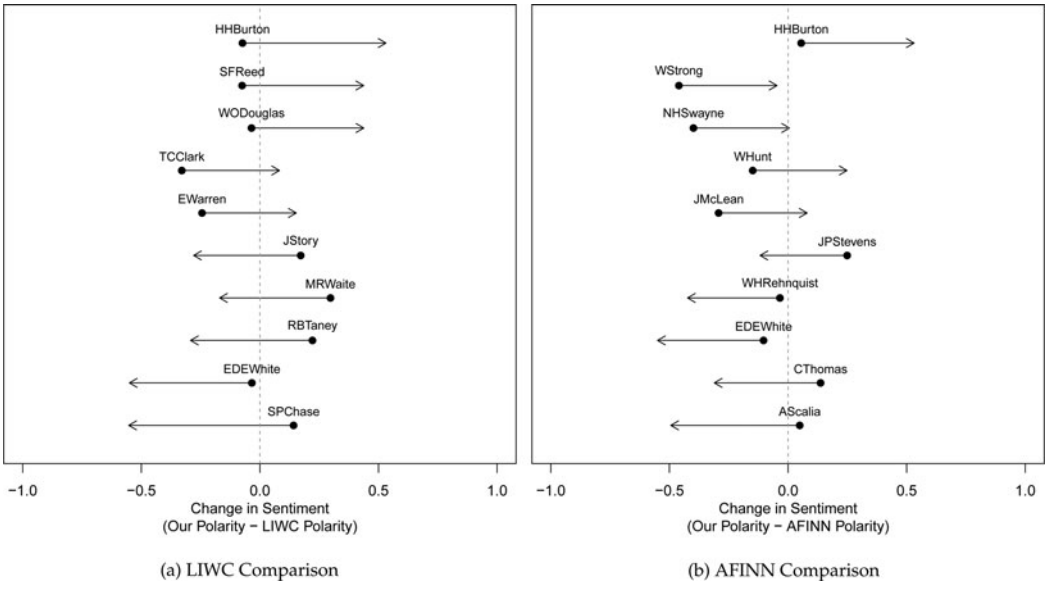


Figure 5. *Major Differences in Polarity Estimates* The above plots the five most positive and five most negative changes in justice-level polarity averages between common dictionary approaches (indicated by black dots) and our estimate (indicated by point of arrow).

across both models; our approach identifies him as the most positive justice on the Court, on average. Historical narratives bear our finding out, as Burton's "affable personality brought together colleagues who sometimes regarded one another with acrimony" and who generally was able to form congenial alliances amongst disparate factions (Rise, 2006, 103). While our approach ranks Burton as the most positive, on average, in his authorship of majority opinions, AFINN ranks Burton 37th of 84 justices, while LIWC ranks Burton 52nd. The Burton example is instructive of the broader changes in justice-level polarity contingent on the choice of dictionary. Looking over the history of the Court, our understanding of where justices stand in terms of their use of emotionally-valenced language is fundamentally re-structured if we look at our estimates as compared to LIWC or AFINN.

Likewise for other justices. Our approach yields substantial drops in the polarity of Chief Justice White's majority opinions. Discussing White's leadership, Associate Justice and later Chief Justice Charles Evans Hughes remarked that his own success as chief stemmed from watching White and learning what *not* to do and the pitfalls one must avoid (Pratt, 1999). For other justices, the results are similarly supportive. We see a substantial drop in the polarity of Chief Justice Morrison Waite's average polarity, from marginally positive in LIWC to marginally negative in our assessment. The Court, during Waite's tenure as chief, was under an enormous workload of which Waite had assigned a substantial portion to himself; moreover, Waite made an emphasis of publicly presenting unity while privately the justices were in disagreement. To wit, his "personal docket books show …[o]f the 247 cases disposed of by the Court during the 1887 term prior to Waite's death, conference dissents were recorded in 35 percent and public dissents in only 10 percent" (Stephenson, 1973, 918). Finally, compared to the LIWC estimates, we find that Chief Justice Roger Taney wrote significantly more negative majority opinions. Such a result is not surprising, as by the time of his death, "Taney was a minority justice, ignored by the president and Congress, held in contempt by the vast majority of his countrymen, and respected only in those places that proclaimed themselves no longer in the Union" (Finkelman, 2006, 540). In all, the alignment of the historical record and the observed shifts in polarity offers suggestive—though not conclusive—support for our approach.

Discussion

Our goal at the outset was to develop a method for building sentiment dictionaries that yield valid, reliable measures of sentiment for corpora of specialized language, and to do so in a way that minimizes the amount of human coding—and associated cost—necessary. Such a method would be very valuable for analyzing text where standard plain-language sentiment dictionaries fail. We characterize our approach as "minimally supervised" (e.g., Uszkoreit et al., 2009) in the sense that it requires a small amount of initial human coding but is largely unsupervised in nature. A natural question is when this "minimally supervised" approach should be selected instead of standard off-the-shelf dictionaries. Given that our results suggest our approach exceeds the performance of these standard dictionaries at least in one area where benchmarks exist, the approach is especially useful in circumstances where language is specialized and/or when its use changes over time. Where specialized language is not expected or specialized dictionaries are already available, our results suggest our approach would perform at least equivalently.

In closing, we note a number of future directions for this research. One key question is the generalizability of our methods: To what extent do our approaches "travel well," yielding valid dictionaries for widely-varying types of specialized vocabularies? One concern on this front has to do with variation in the usage of sentiment-laden words within documents. That is, in the above we have calculated a document-level measure of polarity, but recent work has regularly sought to capture sentiment in shorter portions of texts, including paragraphs, sentences, and phrases. One avenue in which this research must develop is to identify these changes within documents, particularly long-form documents such as Supreme Court opinions. Similarly, the document level measure of polarity obscures a great deal of information on the subjects of the speech. Moving to an analysis of shorter fragments of speech also potentially permits recovering this lost information on the subject of sentiment expression. Finally, the approach itself stands to be upgraded; one clear avenue is to build on the work of Pang and Lee (2004) and to identify and retain only subjectively valenced terms for dictionary construction, removing many potentially noisy terms that undercut classification accuracy. We leave this to future research.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2019.10

References

Black R, Treul S, Johnson T and Goldman J (2011) Emotions, oral arguments, and Supreme Court decision making. *Journal of Politics* **73**, 572–581.

Black R, Hall M, Owens R and Ringsmuth E (2016) The role of emotional language in briefs before the US Supreme Court. *Journal of Law & Courts* **4**, 377–407.

Bryan A and Ringsmuth E (2016) Jeremiad or weapon of words?: the power of emotive language in Supreme Court dissents. *Journal of Law & Courts* **4**, 159–185.

Caldeira G and Zorn C (1998) Of time and consensual norms in the Supreme Court. *American Journal of Political Science* **42**, 874–902.

Danelski D (1960) The influence of the chief justice in the decisional process of the Supreme Court. In *Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois*.

Dave K, Lawrence S and Pennock D (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *12th International World Wide Web Conference*.

Epstein L, Landes W and Posner R (2011) Why (and when) judges dissent: a theoretical and empirical analysis. *Journal of Legal Analysis* **3**, 101–137.

Finkelman P (2006) *Biographical Encyclopedia of the Supreme Court: The Lives and Legal*, Chapter Roger Brook Taney, Washington, DC: CQ Press, pp. 531–541.

Gerner D, Schrodt P, Francisco R and Weddle J (1994) The analysis of political events using machine coded data. *International Studies Quarterly* **38**, 91–119.

Grimmer J and Stewart B (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**, 267–297.

Hansen L, Arvidsson A, Nielsen F, Colleoni E and Etter M (2011) Good friends, bad news—affect and virality in twitter. In *The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet)*.

Haynie S (1992) Leadership and consensus on the U.S. Supreme Court. *Journal of Politics* **54**, 1158–1169.
Hendershot M, Hurwitz M, Lanie D and Pacelle R (2013) Dissensual decision making: revisiting the demise of consensual norms with the U.S. Supreme Court. *Political Research Quarterly* **66**, 467–481.
Liu B (2010) Sentiment analysis and subjectivity. In Indurkya N and Damerau F (eds). *Handbook of Natural Language Processing*, 2nd Edn. Boca Raton, FL: Chapman and Hall/CRC Press, pp. 627–666.
Maas A, Daly R, Pham P, Huang D, Ng A and Potts C (2011) Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.
Mikolov T, Chen K, Corrado G and Dean J (2013a) Efficient estimation of word representation in vector space. In *ICLR Workshop*.
Mikolov T, Sutskever I, Chen K, Corrado G and Dean J (2013b) Distributed representation of words and phrases and their compositionality. In *NIPS*.
Nematzadeh A, Meylan S and Griffiths T (2017) Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.
Nielsen F (2011) A new anew: evaluation of a word list for sentiment analysis in microblogs. In *The ESQ2011 Workshop on "Making Sense of Microposts"*.
Pang B and Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics*, pp. 271–278.
Pang B, Lee L and Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86.
Pennebaker J, Francis M and Booth R (2001) *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers.
Pennington J, Socher R and Manning CD (2014) Glove: global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
Pratt W (1999) *The Supreme Court Under Edward Douglass White, 1910–1921*. Columbia, SC: University of South Carolina Press.
Quinn K, Monroe B, Crespin M, Colaresi M and Radev D (2010) How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* **54**, 209–228.
Rice D (2017) Issue divisions and U.S. Supreme Court decision making. *Journal of Politics* **79**, 210–222.
Rise E (2006) *Biographical Encyclopedia of the Supreme Court: The Lives and Legal*, Chapter Harold Hitz Burton, Washington, DC: CQ Press, pp. 100–104.
Salamone M (2013) Judicial consensus and public opinion: conditional response to Supreme Court majority size. *Political Research Quarterly* **67**, 320–334.
Selivanov D (2016) *text2vec: Modern Text Mining Framework for R*. R package version 0.4.0.
Spaeth HJ, Epstein L, Ruger TW, Whittington KE, Segal JA and Martin AD (2012) The Supreme Court database. http://supremecourtdatabase.org.
Stephenson DG (1973) The chief justice as leader: the case of morrison waite. *William and Mary Law Review* **14**, 899–927.
Tang D, Wei F, Yang N, Zhou M, Liu T and Qin B (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, pp. 1555–1565. Association for Computational Linguistics.
Tang D, Wei F, Qin B, Yang N, Liu T and Zhou M (2016) Sentiment embeddings with applications to sentiment analysis. *Knowledge and Data Engineering, IEEE Transactions on* **28**, 496–509.
Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424.
Uszkoreit H, Xu F and Li H (2009) Analysis and improvement of minimally supervised machine learning for relation extraction. In *NLDB09 Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems*.
Walker T, Epstein L and Dixon W (1988) On the mysterious demise of consensual norms in the United States Supreme Court. *Journal of Politics* **50**, 361–389.
Wang P and Domeniconi C (2008) Building semantic kernels for text classification using wikipedia. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 713–721.
Zink J, Spriggs J and Scott J (2009) Courting the public: the influence of decision attributes on individuals' views of court opinions. *Journal of Politics* **71**, 909–925.

Cite this article: Rice DR, Zorn C (2019). Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods* 1–16. https://doi.org/10.1017/psrm.2019.10