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

March 27, 2017

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 U.S. federal appellate court decisions.

Word Count: 8,656

[*] On publication, all materials necessary to replicate the results reported herein will be posted to the *AJPS* Dataverse.

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, Lee and Vaithyanathan, 2002; Pang and Lee, 2004; Wilson, Wiebe and Hoffmann, 2005; Maas et al., 2011) rely on classifying or scoring 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 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.

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, Francis and Booth, 2001; Pennebaker et al., 2007), and many have already found their way into published work (e.g., Black et al., 2011; Bryan and Ringsmuth, 2016; Black et al., 2016). Pre-constructed dictionaries thus offer superlative ease of use. But while they have been applied across a variety of contexts, it is also the case that 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; Quinn et al., 2010), 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 (that is, 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 U.S. 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 alternative for extracting emotional context from text. At the same time, dictionary-based approaches have many limitations. Pre-constructed dictionaries for use with modern standard U.S. 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 reflect the emotional valence of the language as it is used in context, and so are more likely to yield accurate estimates of sentiment in specialized vocabularies. Such dictionaries, however, are also difficult, expensive, and time-consuming to construct, since they typically involve specifying every emotionally-valenced word or phrase that could be encountered 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 general approach to building specialized sentiment dictionaries leverages both the structure of language and the corpus of text itself. That is, it builds a dictionary from the words used in the texts from which sentiment is to be extracted, and does so by relying on some universal facts about how words are used.[1] Specifically, as our goal is to select a set of words related to a particular dimension of interest (here, sentiment), we seek to automatically identify the 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 solely based on identifying words which occur in similar contexts within the corpus.

Importantly, extensive prior work in natural language processing has focused on automatically identifying semantically similar words. Building on this work, researchers

[1] For simplicity, we focus on the simplest form of sentiment analysis, the extraction of positive or negative sentiment.

have recently sought to identify methods for understanding a word's embedding, or context within a particular word space. Identifying the appropriate and relevant word space, however, is increasingly difficult and complicated as the amount of unstructured textual data increases. Recent work by Mikolov, Chen, Corrado and Dean (2013) proposed the Skip-gram model, based on a shallow neural network, for identifying word representations "useful for predicting the surrounding words in a sentence or a document" (2). 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, Sutskever, Chen, Corrado and Dean, 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`). A host of methods and extensions have been proposed through which to perform these calculations (Mikolov, Sutskever, Chen, Corrado and Dean, 2013, e.g.,); we utilize the `text2vec` implementation of `GloVe` (Pennington, Socher and Manning, 2014) available in R.

It is this property – semantic similarity based on similar word usage – which we leverage to automatically construct dictionaries. Vector similarity is evaluated by cosine similarity, consistent with prior work (e.g., Maas et al., 2011). 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.[2] These new words – semantically similar to the uncontroversially positive and negative seed words *within the domain*

[2] 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 "incor-

– are then extracted in order to construct the dictionary.[3]

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 counts of these terms, the process also provides a wealth of important information on the terms; specifically, we know the distribution of term usage across the corpus (term frequency - inverse document frequency [tf-idf]) and the similarity of the term's vector representation to the positive (negative) vector. Making use of these pieces of 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:[4]

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

We think of this approach as "minimally supervised," in that it resembles in most respects unsupervised / data-driven approaches (e.g., clustering) but requires at the outset a very small amount of human selection of the seed sets to serve as starting points for the learn-

rect." Note that it would be possible to increase classification accuracy by identifying highly informative terms for the context (for example, "recommend" in the context of movie reviews.

[3] 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.

[4] For purposes of comparison, we center and scale the polarity scores. As will be noticeable in Figure 1, this has a negligible impact on the accuracy of our approach but substantially improves the accuracy of both AFINN and LIWC.

ing algorithms. Though the improvements over standard off-the-shelf dictionaries are clear, our approach also 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. That is, 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 monumentally faster than the alternatives; the decrease in monetary 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.[5] These data, introduced in Maas et al. (2011), consist of 100,000 movie reviews – 25,000 positive, 25,000 negative, and 50,000 unlabeled – extracted from the Internet Movie Database (IMDB) archive. The assignment of positive or negative codes for these reviews is explicitly based on ratings provided by the reviewers. Prior research has utilized these or similar ratings and text extensively, primarily in the development of machine learning methods for the identification and measurement of sentiment in texts (e.g., Pang, Lee and Vaithyanathan, 2002; Wang and Domeniconi, 2008; Maas et al., 2011; Dinu and Iuga, 2012). For our purposes,

[5] These data are available online at `http://ai.stanford.edu/~amaas/data/sentiment/`

the assigned positive and negative ratings in the Movie Review Data provide a benchmark sentiment dataset by which we can assess the validity of our approach. 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, Lawrence and Pennock, 2003; Pang and Lee, 2004). 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 skip-gram window of 50 tokens. To estimate the model, we use 300-dimensional vectors, setting $X_{max} = 10$.

We assess the vector representations through an analogies test. To benchmark the semantic validity of estimated vectors, researchers developed the analogies task similar to the above example – vec(`woman`) + vec(`king`) - vec(`man`) – but on a much larger scale. Specifically, the analogies task features 19,544 analogies covering both syntactic and semantic content (Pennington, Socher and Manning, 2014). The task itself is demanding; only the *most* similar vector is identified as a correct match. Though demanding, the task provides an immediate tool for assessing whether semantically similar terms have been identified. Of the 19,544 possible questions, we have each of the four tokens for 4,323; of those 4,323, we correctly predict the token in 33.54% of tasks. Among sub-categories, our accuracy ranges from a low of 0 (currency, city-in-state) to a high of 80% (commonly mentioned country capitals). Two facts caution against concluding the model is poorly fit; first, the corpus is much smaller than those frequently employed elsewhere. As a comparison, we have 198,409 tokens taken from a specialized corpus, whereas the corpora analyzed in Pennington, Socher and Manning (2014) range from 1 to 42 billion tokens and are more general (e.g., Wikipedia). Learning vector representations in smaller, specialized settings is thus more difficult. Likewise, recall that the task itself is highly demanding; only the *most* similar word counts as a match, meaning a miss ranking the word second-most similar and least similar are treated identically. The vec(`woman`) + vec(`king`) - vec(`man`) task is instructive; the five most similar vectors are, in decreasing order of similarity, "king", "queen", "kong", "princess", and "stephen." The results make clear both the specialized relations of terms within the corpus ("king kong", "stephen king") while also demonstrating the difficulty of the analogies task as it would be treated as a wrong answer. We explore below the influence of variation in both of these parameters (corpus size and analogies accuracy) on measurement validity.

We extract the top 500 positive and top 500 negative words by cosine similarity and calculate polarity according to the description above.[6] 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 12,004 reviews as negative and 12,663 as positive, with an overall classification accuracy of 80.4%.[7] For purposes of illustration and comparison, in Figure 1we plot the estimated polarity of different dictionary-based approaches (x-axis) against the assigned ratings (y-axis), with a lowess line demonstrating fit. Overall it is clear that our approach is generally performing well, with the lowess 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 zero cutpoint, with reviews in this region frequently referencing

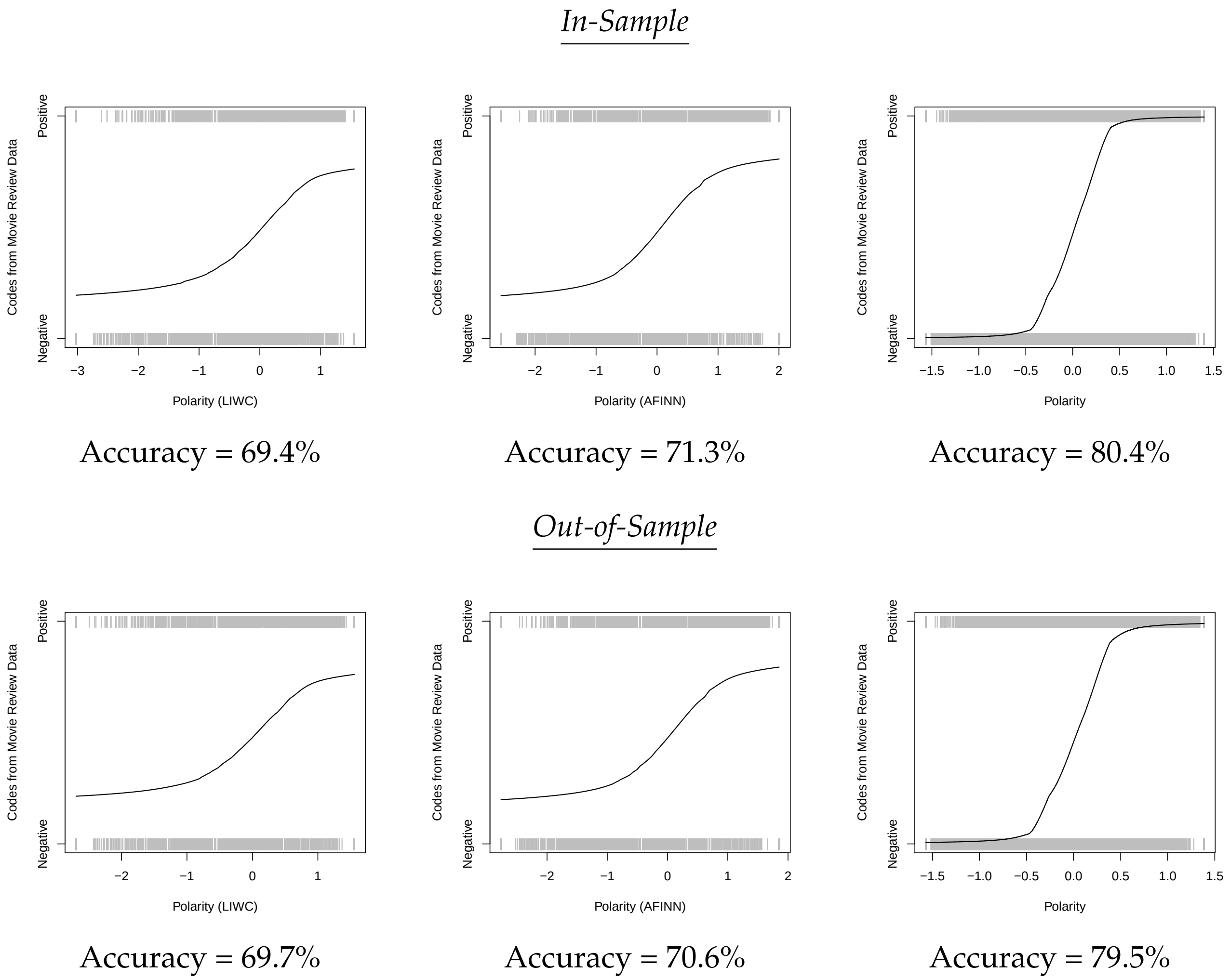
[6] We explore the influence of variation in the word cutoff later. Adding additional words does little other than contribute noise to the estimates.

[7] The remaining 332 are classified as neutral as they feature no positive or negative words.

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 fee-based Linguistic Inquiry and Word Count (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). In work comparing the results from AFINN to results obtained from other pre-specified dictionaries, AFINN rated behind only SentiStrength (Nielsen, 2011). 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. Lowess 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*.



In-Sample

Accuracy = 69.4%

Accuracy = 71.3%

Accuracy = 80.4%

Out-of-Sample

Accuracy = 69.7%

Accuracy = 70.6%

Accuracy = 79.5%

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.

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 two dimensions of our research problem: the size of the corpora and the size of the extracted dictionary. Though minimally supervised, these two dimensions remain at the discretion of the researcher, and thus bear additional discussion.

We begin by assessing accuracy across the size of the corpus used to estimate the word vectors. To do so, we use parameters consistent with those employed in the analysis above.[8] 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

[8] We only shift the minimum number of word occurrences necessary for inclusion by a common ratio across iterations.

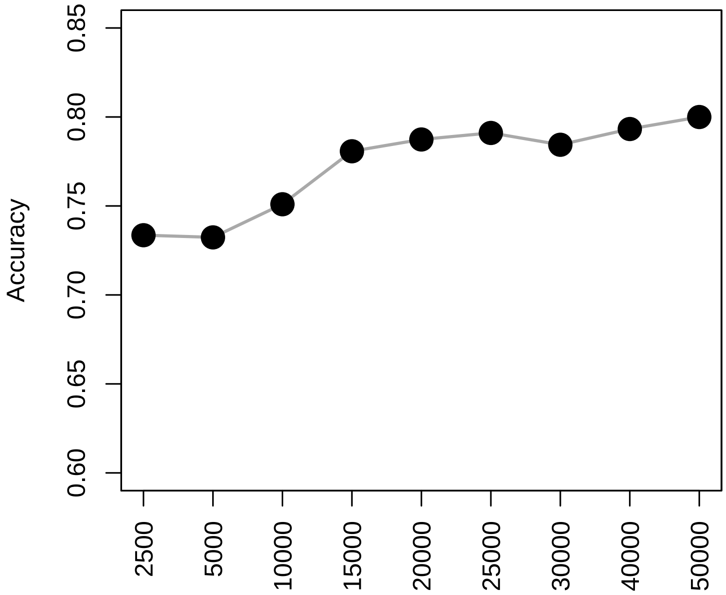


Figure 2: *Robustness To Differences in Corpus Size.* Plot of the accuracy (y-axis) of our polarity approach across variation in the corpus size (x-axis).

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 assessed within each sample, with the neutral reviews excluded in computing accuracy.

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 (2,500 or 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, 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. Third, also notable is that accuracy plateaus from approximately 15,000 documents to 30,000 documents, with moderate improvements thereafter. This plateau provides evidence that the estimates of semantic similarity quickly converge to levels of accuracy that offer substantial improvements on existing dictionary-based methods. 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 converge, 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.

We move next to assessing the accuracy of the classifier across the number of sentiment words identified and stored in the dictionary. Here, 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. Thus, 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 3. Most evident in the plot is the striking lack of variation across the size of the dictionary, with the standard deviation of the entire series standing at 0.7%. Moreover, classification accuracy across the entire series is universally well 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

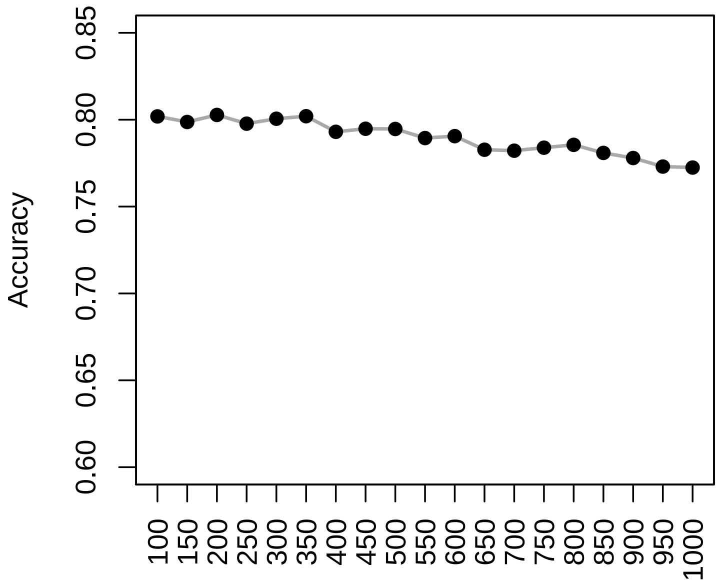


Figure 3: *Robustness To Differences in Size of Extracted Dictionary.* Plot of the accuracy (y-axis) of our polarity approach across variation in the size of the extracted dictionary (x-axis).

the choice of the number of words to extract is of little consequence.

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 financial, 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 points, the former – the accuracy of our approach relative to supervised learning implementations – deserves attention. To demonstrate, in Table 1 we compare our accuracy across a series of seminal works in supervised learning for sentiment analysis of movie review data. In Pang, Lee and Vaithyanathan (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 incredibly 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,

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

Model	**Mean**	**Min**	**Max**
Pang, Lee and Vaithyanathan (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	80.4	-	-

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. Therefore, having documented the validity of our approach, we turn next to a unique domain: the Supreme Court of the United States.

Illustration: Sentiment in U.S. Supreme Court Opinions

In the study of the U.S. 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, Spriggs and Scott, 2009), and on legal development (Epstein, Landes and Posner, 2011; Rice, 2016). 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 – efforts on the part of the Court to hide private disagreement from the public – and the role of the chief justice in precipitating that decline (e.g., Danelski, 1960). Throughout, 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, Epstein and Dixon, 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 themselves. Opinion language is the central mechanism by which the 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 emotional attitudes toward the decisions at hand. Consider *Moran v. Burbine*[9] (1986), which dealt with the Fifth and Sixth Amendment rights of the criminally accused. Writing for the majority, Justice Sandra Day O'Connor rejected Burbine's argument by stating that it 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 method of reaching that conclusion, "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 to the dissent, O'Connor's majority opinion stated that "JUSTICE STEVENS' apocalyptic

[9] 475 U.S. 412.

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 this and many other opinions make clear, 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. Off-the-shelf dictionaries are difficult to apply and validate within this particular domain given the the particular, and at times peculiar, legal language used in them. Alternatively, machine learning options – coding training data or using proxy measures – are costly, difficult to apply, and likely to yield inaccuracies. In the former case, coding training data is complicated by the length of judicial opinions, with the median length of majority opinions approaching 5,000 words in recent terms (Black and Spriggs, 2008). Moreover, using the language of the case syllabus, a shortened description of the case often relied upon in the coding of Supreme Court cases (e.g., Spaeth et al., 2012), is exceedingly unlikely to provide an accurate signal of the opinion's sentiment.

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. For this analysis, we retain only majority opinions. 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.[10] 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 particularly vexing problem for text-as-data classification across long periods of time – as here – is the issue of semantic change. That is, one might reasonably be worried that words which carry a negative valence in 2000 may not carry 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 addressing this semantic shift. 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.[11]

We applied our approach to the resulting bodies of text to estimate the aggregate sentiment of each opinion, and in addition generated sentiment scores for each opinion using the standard pre-constructed LIWC and AFINN sentiment dictionaries. We begin by as-

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

[11] While we base our splits on theoretically motivated subsets of the Court's history, it would also be possible to derive sentiment dictionaries and estimate textual sentiment across a variety of alternative cutpoints. Consider, for instance, computing sentiment across rolling windows of time and averaging the resultant scores while also estimating measures of uncertainty. As we say, the approach is highly flexible.

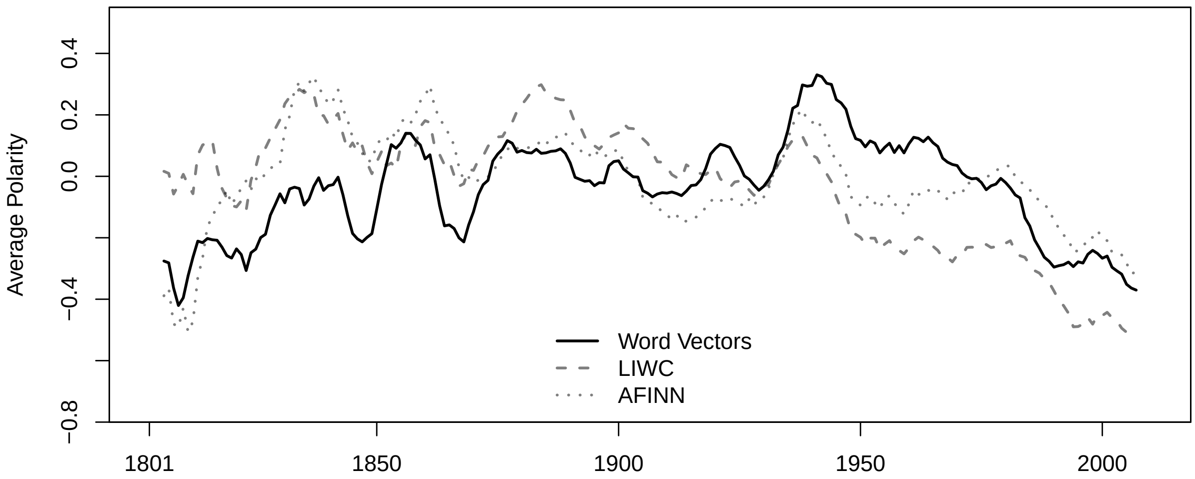
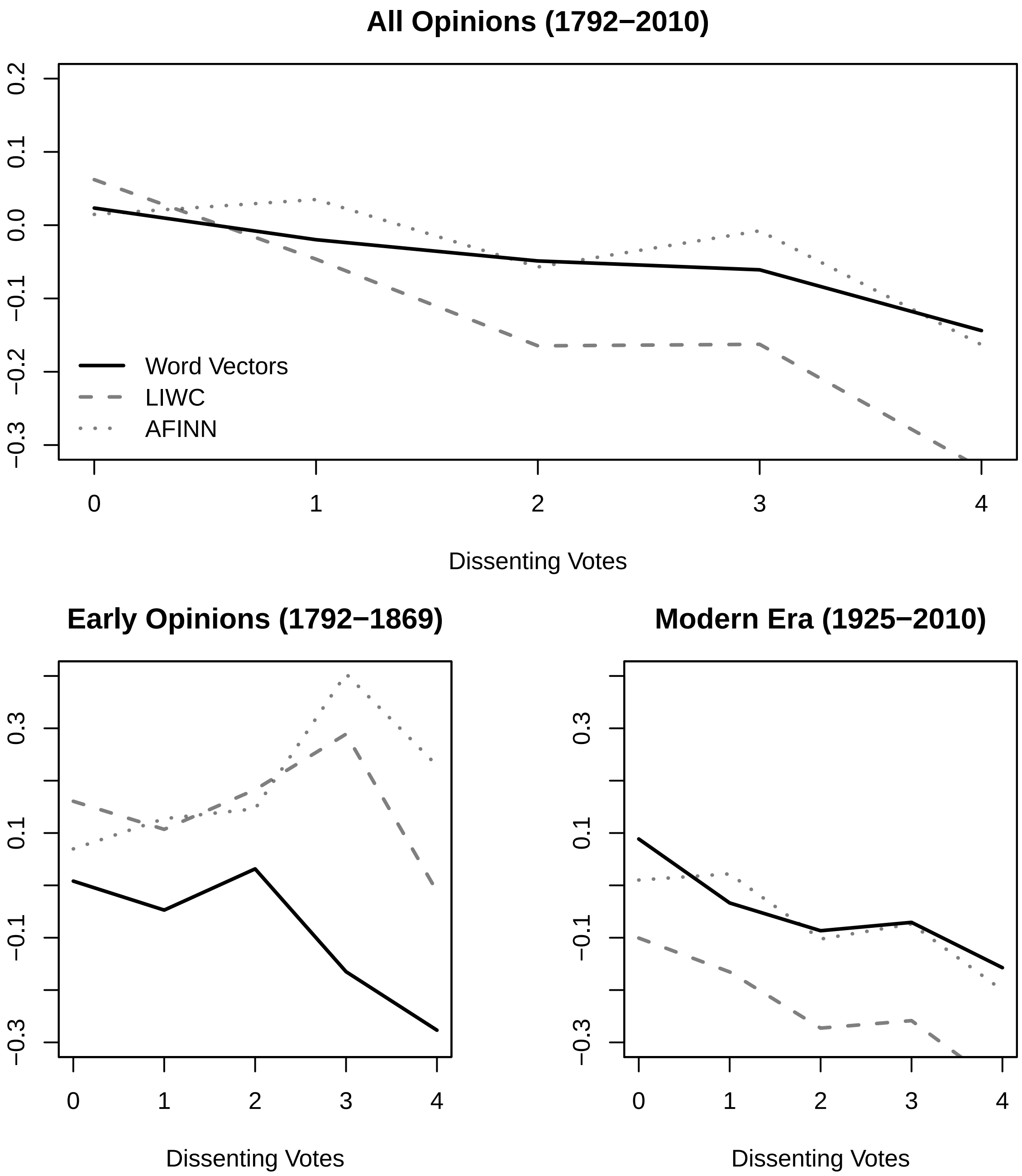


Figure 4: *Seven-year moving average of opinion polarity by year.* Plot of a seven-year moving average (symmetric) of average opinion polarity calculated using our approach (solid black line), the Linguistic Inquiry and Word Count dictionary (long dashed grey line), and the AFINN dictionary (short dashed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

sessing the associations between these measures. Figure 4 shows the seven-year moving average of majority opinion polarity as estimated using each of the three dictionary approaches. While the trends of each line generally mirror one another, the actual values are consistently different. 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.

While informative, Figure 4 does not conclusively demonstrate the superiority of our approach, but rather demonstrates differences and similarities in the resulting trends we might identify. Turning then to validation, we approach the relative performance of each across two criteria: first, the degree to which they correlate with other variables they should theoretically be correlated (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 research frequently employed voting divisions as evidence of the social environment; though imperfect for reasons outlined above, the measure offers a very coarse picture of the Court's level of disagreement and thus we might properly expect that increases in divisiveness would necessarily correlate with more negative majority opinions. To see this, in Figure 5 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 that only our measure documents a consistent decline in polarity by the number of dissenting votes. Both LIWC and AFINN dictionaries, on the other hand, identify increases in majority opinion polarity at various point, most notably in 6 to 3 decisions of the Court. The differences in approaches are most stark, however, in the earliest years of the Court. There, both AFINN and LIWC document *increases* in majority opinion polarity as the number of dissenting votes goes up, with drops only occurring when there are four dissenting votes. On the contrary, our approach yields estimates that – while there is a marginal increase at two dissenting votes which takes the average back to approximately neutral – generally fall in the area of 0 until dropping off after two dissenting votes into strongly negative polarity. Importantly, in the modern era – when standard dictionaries should work best though still suboptimally – our approach yields estimates that are generally correlated but largely more sensible. Consider, for instance, that ours is the only approach that identifies



All Opinions (1792−2010)

Early Opinions (1792−1869)

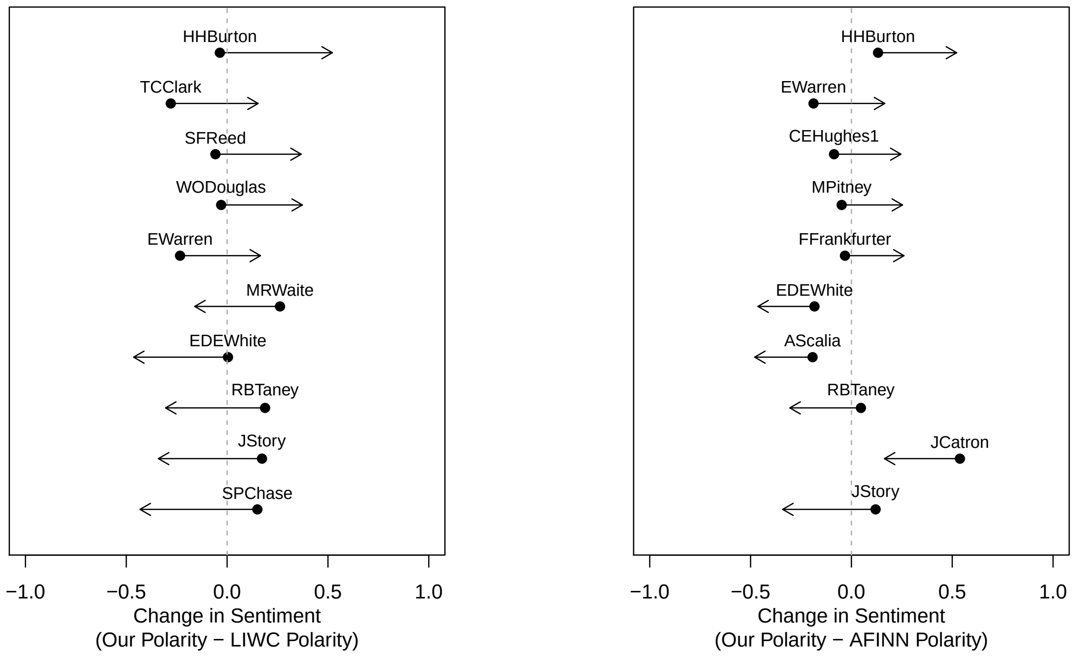
Modern Era (1925−2010)

Figure 5: *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 grey line), and the AFINN dictionary (short dashed gray line).

unanimous opinions as generally positive in tone during the modern era.

With evidence of convergent validity, we move next to examining discriminant validity. To see this, 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 6, 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. 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 across both models; though both AFINN and LIWC regard Burton as approximately neutral, our approach identifies him as one of the more positive justices on the Court. 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).

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



(a) LIWC Comparison

(b) AFINN Comparison

Figure 6: *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).

to Waite's death, conference dissents were recorded in 35 percent and public dissents in only 10 percent" (Stephenson, 1973, 918). Finally, compared to both AFINN and 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 shifts in polarity we observe across individual justices are consistently supported by the historical record, providing further evidence of the validity of 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, Xu and Li, 2009) in the sense that it requires a small amount of initial human coding but is largely unsupervised in nature. Our work here indicates that such an approach performs far better than standard dictionary-based approaches in recovering sentiment, and achieves levels of accuracy at or exceeding standard supervised learning approaches. Moreover, the approach may be especially useful in circumstances where language is specialized and/or when its use changes over time.

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 sentimental 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.

References

Black, Ryan and James Spriggs. 2008. "An Empirical Analysis of the Length of U.S. Supreme Court Opinions." *Houston Law Review* 45:621–682.

Black, Ryan, Matthew Hall, Ryan Owens and Eve Ringsmuth. 2016. "The Role of Emotional Language in Briefs before the US Supreme Court." *Journal of Law & Courts* 4(2):377–407.

Black, Ryan, Sarah Treul, Timothy Johnson and Jerry Goldman. 2011. "Emotions, Oral Arguments, and Supreme Court Decision Making." *Journal of Politics* 73(2):572–581.

Bryan, Amanda and Eve Ringsmuth. 2016. "Jeremiad or Weapon of Words?: The Power of Emotive Language in Supreme Court Dissents." *Journal of Law & Courts* 4(1):159–185.

Caldeira, Gregory and Christopher Zorn. 1998. "Of Time and Consensual Norms in the Supreme Court." *American Journal of Political Science* 42:874–902.

Danelski, David. 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, Kushal, Steve Lawrence and David Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *12th International World Wide Web Conference*.

Dinu, Liviu and Iulia Iuga. 2012. "The Naive Bayes Classifier in Opinion Mining: In Search of the Best Feature Set." *Computational Linguistics and Intelligent Text Processing* 7181:556–567.

Epstein, Lee, William Landes and Richard Posner. 2011. "Why (And When) Judges Dissent: A Theoretical and Empirical Analysis." *Journal of Legal Analysis* 3(1):101–137.

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

Gerner, Deborah, Philip Schrodt, Ronald Francisco and Judith Weddle. 1994. "The Analysis of Political Events using Machine Coded Data." *International Studies Quarterly* 38:91–119.

Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:forthcoming.

Hansen, Lars, Adam Arvidsson, Finn Nielsen, Elanor Colleoni and Michael Etter. 2011. Good Friends, Bad News - Affect and Virality in Twitter. In *The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet)*.

Haynie, Stacia. 1992. "Leadership and Consensus on the U.S. Supreme Court." *Journal of Politics* 54:1158–1169.

Hendershot, Marcus, Mark Hurwitz, Drew Lanie and Richard Pacelle. 2013. "Dissensual Decision Making: Revisiting the Demise of Consensual Norms with the U.S. Supreme Court." *Political Research Quarterly* 66(2):467–481.

Liu, Bing. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, ed. Nitin Indurkya and Fred Damerau. Chapman and Hall/ CRC Press pp. 627–666.

Maas, Andrew, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Distributed Representation of Words and Phrases and their Compositionality. In *NIPS*.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient Estimation of Word RRepresentation in Vector Space. In *ICLR Workshop*.

Nielsen, Finn. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *The ESQ2011 Workshop on 'Making Sense of Microposts'*.

Pang, Bo and Lillian Lee. 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, Bo, Lillian Lee and Shivakumar Vaithyanathan. 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, James, Cindy Chung, Molly Ireland, Amy Gonzales and Roger Booth. 2007. *The Development and Psychometric Properties of LIWC2007*. Austin, TX: LIWC.
URL: *www.liwc.net*

Pennebaker, James, Martha Francis and Roger Booth. 2001. *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers.

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.
URL: *http://www.aclweb.org/anthology/D14-1162*

Pratt, Walter. 1999. *The Supreme Court Under Edward Douglass White, 1910-1921*. University of South Carolina Press.

Quinn, Kevin, Burt Monroe, Michael Crespin, Michael Colaresi and Dragomir Radev. 2010. "How to Analyze Political Attention With Minimal Assumptions and Costs." *American Journal of Political Science* 54:209–228.

Rice, Douglas. 2016. "Issue Divisions and U.S. Supreme Court Decision Making." *Journal of Politics* .

Rise, Eric. 2006. *Biographical Encyclopedia of the Supreme Court: The Lives and Legal*. Washington, DC: CQ Press chapter Harold Hitz Burton, pp. 100–104.

Salamone, Michael. 2013. "Judicial Consensus and Public Opinion: Conditional Response to Supreme Court Majority Size." *Political Research Quarterly* 67(2).

Spaeth, Harold J., Lee Epstein, Theodore W. Ruger, Keith E. Whittington, Jeffrey A. Segal and Andrew D. Martin. 2012. "The Supreme Court Database." http://supremecourtdatabase.org.

Stephenson, D. Grier. 1973. "The Chief Justice As Leader: The Case of Morrison Waite." *William and Mary Law Review* 14(4):899–927.

Turney, Peter. 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, Hans, Feiyu Xu and Hong Li. 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*. pp. 8–23.

Walker, Thomas, Lee Epstein and William Dixon. 1988. "On the Mysterious Demise of Consensual Norms in the United States Supreme Court." *Journal of Politics* 50:361–389.

Wang, Pu and Carlotta Domeniconi. 2008. Building semantic kernels for text classification using wikipedia. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 713–721.

Wilson, Theresa, Janyce Wiebe and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 347–354.

Zink, James, James Spriggs and John Scott. 2009. "Courting the Public: The Influence of Decision Attributes on Individuals' Views of Court Opinions." *Journal of Politics* 71(3).