Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies*

Douglas R. Rice
Department of Political Science
University of Mississippi
drrice@olemiss.edu

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

Version 0.4

February 12, 2015

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 develop a class 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 through comparison to a well-known standard (nonspecialized) sentiment dictionary, compare its performance to (costlier) supervised approaches, and show its usefulness in an application to the specialized language used in U.S. federal appellate court decisions.

[*] A previous version of this paper was presented at the Annual Meeting of the Midwest Political Science Association, April 3-6, 2014, Chicago, Illinois. The paper remains preliminary; the content is likely to change with future revisions. Please be aware of this if you choose to cite this version.

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) and opinion content. 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) 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 ("test") 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, and 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 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 preconstructed dictionaries are publicly available, such as the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker, Francis and Booth, 2001; Pennebaker et al., 2007). 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 a set of approaches 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 necessary to reliably create a valid dictionary). The next section of the paper describes the problem in general terms, and outlines our proposed method. In the third section, we validate our method through its application to a well-understood corpus of data on film reviews, and compare our results with those based on standard supervised learning approaches. The fourth section applies our approach to a specialized context, the use of language in U.S. federal appellate (and, specifically, Supreme Court) opinions, as a means of measuring the degree of interpersonal harmony on the Court. The section not only demonstrates the utility of our approach in generating measures of substantive interest to social scientists, but also validates our approach as a superior alternative to utilizing reasonable, but sub-optimal, training data in supervised learning classification. Section five concludes.

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, both types of dictionary-based approaches offer potential limitations as well. 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 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.

Co-Occurrence and Conjunction

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. For simplicity, we focus on the simplest form of sentiment analysis,t he extraction of positive or negative sentiment. At this writing, our general method encompasses two specific implementations. In both instances, we begin with two sets of "seed words" – one positive, one negative – specified by the analyst. These are comprised of a relatively small number of commonly-used words (perhaps 20 or 25)[1] which are universally positive (/ negative) irrespective of their context.

[1] In the examples described below, we initially chose 25 positive and 25 negative seed words.

We then adopt two separate techniques – word *conjunction* and word *co-occurrence* – to create dictionaries using only the actual texts of the opinions and our small seed sets of positive and negative terms.

The simpler of our two methods we term our *conjunction* approach, where we adapt the insights of Hatzivassiloglou and McKeown (2007) to construct our dictionary. We begin by identifying words in each text which are explicitly *conjoined* with words in our seed sets. For example, if the word stem `delight*` was among our initial set of positive seed words, and a text contained the sentence "I am honored and delighted to be speaking to you today," then we would say that `delight*` and `honor*` were conjoined. For each text in our corpus, we identify the presence of our seed words and whether those seed words are used in conjunctions. All words used in conjunctions with our one of the words in our seed sets are then retained for inclusion in the respective (positive or negative) dictionary. The resulting positive and negative dictionaries are then cleaned of irrelevant and inappropriate terms, yielding a conjoined dictionary of context-specific valenced terms derived directly from the texts.

The *co-occurrence* approach is similar, but relies instead on identifying words which co-occur in texts with the seed words. A word is said to co-occur with a seed word if the two words both appear anywhere in the same text. After identifying all incidents of co-occurrence with each of our seed words, we build a co-occurrence matrix of our seed sets and the remaining terms, with each entry in the matrix representing a count of the number of texts in which the two terms both appeared (i.e., co-occurred). More specifically, for each non-seed word i in the corpus, we calculate $k_{i(p)}$, the proportion of times it co-occurred in a document with one of the positive seeds, and $k_{i(n)}$, the proportion of co-occurrence with one of the negative seeds. From these counts, we calculate the odds of word co-occurrence with positive seeds as $O_p = \frac{k_{i(p)}}{1-k_{i(p)}}$ and those of co-occurrence with negative seeds as $O_n = \frac{k_{i(n)}}{1-k_{i(n)}}$. We then estimated the log odds ratio as $\ln \left[\frac{O_p}{O_n}\right]$

and multiply this by the total number of co-occurrences to arrive at an estimate of word polarity. Finally, depending on the size of the corpus and the total number of unique terms within the corpus, we retain some number of positive and negative terms based on the value of our word polarity scores.

In both instances, the result is a pair of sentiment dictionaries – one comprised of positive words, one of negative terms – that is derived from, and specific to, the corpus of text being analyzed. These dictionaries can then be used individually or in an ensemble to rate the sentiment of the texts in question. The result is an approach that we think of 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 coding of the seed sets to serve as starting points for the learning algorithms.

Practical Implementation

As a practical matter, our methods require a degree of preprocessing of data to be effective. To prepare the texts for analysis, we remove all punctuation and capitalization; while these may be informative of sentiment in certain applications (notably, social media), they are unlikely to be informative in the applications we contemplate. Similarly, because we are interested in estimating sentiment, we also address negations, which can invert the polarity of words. Consistent with prior research (Das and Chen, 2007; Pang, Lee and Vaithyanathan, 2002), we prefixed negation terms ("not," "no") to all subsequent terms until the next punctuation mark. For instance, the phrase "not competent" in a text would be altered to "not-competent." We then used the Stanford Part-of-Speech Tagger (Toutanova et al., 2003) to tag word types in all opinions, and extracted only those terms which potentially had a positive or negative valence.[2] Finally, in addition to our data-based dictionaries, we also estimate polarity using the LIWC software. This yields an

[2] Specifically, we retained only adjectives, adverbs, and nouns.

estimate consistent with dictionaries employed in the measurement of similar concepts in previous political science research (Owens and Wedeking, 2011, 2012) but also potentially biased by the inclusion of contextually inappropriate terms.

Our trio of approaches yields three dictionaries – LIWC, co-occurrence, and conjoined – of positively and negatively valenced terms. With the dictionaries in hand, we calculate a simple, dictionary-specific measure of text polarity as:

$$\text{Polarity}_i = \frac{N \text{ of Positive Words}_i - N \text{ of Negative Words}_i}{N \text{ of Positive Words}_i + N \text{ of Negative Words}_i}$$

As we discuss below, in practice we suggest averaging across these dictionary-specific estimates; an important question for this and future work is the value of doing so.

Validation: Movie Review Data

We begin by testing our approach with the Cornell Movie Review Data.[3] These data, introduced in (Pang and Lee, 2004), consist of 2000 movie reviews – 1000 positive and 1000 negative – 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 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; Dinu and Iuga, 2012). For our purposes, the assigned positive and negative ratings in the Movie Review Data provide a benchmark sentiment dataset by which we can assess the validity our approach. An added benefit is derived from the fact that the sentiment of

[3] These data are available online at `www.cs.cornell.edu/people/pabo/movie-review-data/`.

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.

As detailed above, for our approach we estimate our three measures of polarity, then average across approaches to estimate document polarity. We therefore begin by estimating sentiment using LIWC's pre-determined dictionary. In the upper-left plot in Figure 1, we have plotted LIWC's estimated polarity against the assigned positive and negative ratings. This particular estimate provides 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 LIWC polarity measure, it classifies just 58.5% of cases correctly. While better than the baseline of 50%, LIWC disproportionately assigns positive ratings to movie reviews. To wit, 78% of cases are classified as positive, a fact evident in the plot in Figure 1. Though there are potentially multiple reasons for this, one is that LIWC does not address negations, which can invert the polarity of terms.

Therefore, we move to dictionaries created from the conjoined and co-occurrence approaches. Recall that, as outlined above, we address negations in each of these approaches by adding a prefix to all terms subsequent to negation terms ("no", "not", etc.), until we encounter a punctuation mark. After doing so, we identify a seed set of positive and negative terms appropriate for the context of movie reviews. To create the conjoined dictionary, we then run the program to extract new terms conjoined to words in our seed sets. Having extracted the terms, we stem the terms and remove all duplicates. The resultant conjoined dictionary has 163 total terms, 78 of which are positive and 85 of which are negative. To create the co-occurrence dictionary, we follow the steps outlined above and identify the words which are most likely to co-occur with our seed set within the corpus. We extract the top 200 positive and top 200 negative words, add the seed words to the dictionary, and remove a list of common stop words, yielding a co-occurrence dictionary of 437 terms, 216 of which are positive and 221 of which are negative. From each dictionary, we then estimate polarity. Again, the results are presented in Figure 1. The figure provides evidence that both new dictionaries, derived from the texts using a set of seed words, do not have the same bias towards positive coding that was evident with LIWC.

Finally, we plot the average of the three measures of polarity in the lower-right corner of Figure 1. In all, the accuracy of our measure is 72.5 percent. While far from perfect, these results come close to matching the reported accuracies of machine learning approaches with the Movie Review Data (Pang, Lee and Vaithyanathan, 2002). In their work, Pang, Lee and Vaithyanathan (2002) report the results using different feature sets (unigrams, bigrams, etc.) across three classifiers – naive Bayes, maximum entropy, and support vector machines – with classification accuracy ranging between 72.8% and 82.9%. With an accuracy rate of 72.5%, our approach comes close to matching the accuracy of machine learning classifiers which, it should be noted, are explicitly trained to optimize predictive accuracy.

Note here that we do not argue that our approach is a substitute for these or any other supervised machine learning procedure. 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, the above demonstrates that our approach yields estimates close to the best-performing machine learning classifiers. Therefore, having documented the validity of our approach, we turn next to one such research area with dubious, though plausible, benchmarks: the Supreme Court of the United States.

Application: Sentiment in U.S. Supreme Court Opinions

More than a half-century ago, David Danelski authored what might well be the most important unpublished conference paper in the field of judicial politics.[4] Danelski's "The Influence of the Chief Justice in the Decisional Process of the Supreme Court" (Danelski, 1960) spawned decades of research, much of it focused on the ability of the Chief Justice of the U.S. Supreme Court to influence the degree of consensus on the Court. A central challenge in this research 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 make it likely that such indicators will mask the true level of disagreement on the Court.

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*[5] (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 un-

[4] This section is based in part on Rice and Zorn (2013).

[5] 475 U.S. 412.

tenable," 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 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 (Spaeth et al., 2012, e.g.,), is exceedingly unlikely to provide an accurate signal of the opinion's sentiment. The latter option, training a classifier based on previously utilized measures such as voting patterns (e.g., Corley et al., 2010; Corley, Steigerwalt and Ward, 2013) or dissenting and concurring behavior (e.g., Walker, Epstein and Dixon, 1988; Caldeira and Zorn, 1998; Hendershot et al., 2013), carries perhaps greater appeal. These previously utilized metrics are, certainly, reflective of divisions of opinion on the Court. A classifier trained to predict these divisions using solely the language of opinions could potentially differentiate "strong" and "weak" opinions, for instance. More practically, it could be used to classify all of the more than 150 years of opinions not included in the Supreme Court database. However, to the extent these dynamics are not tightly-connected to the concept of interest (here, sentiment) they introduce potentially serious biases in the analyses. We turn now to evaluating one such approach, and offer evidence that ours, based on building a sentiment dictionary directly from the domain, offers a superior alternative.

Supervised Learning

Recall that, to identify disagreement on the Court, prior quantitative research has generally relied either on the authoring of concurring and dissenting opinions or on voting divisions on the merits of each case. That research thus suggests two diametrically opposing indicators of the Court's social environment: on the negative side, we have dissenting opinion(s), while on the positive side, we have unanimous majority opinions. With that in mind, we treat dissenting opinions as "negative" and unanimous (9-0) majority opinions as "positive," and use these to train a supervised classifier. We can then treat the estimated probability of the case being a majority opinion as a continuous measure of the polarity of the opinion; higher values indicate the classifier assigns higher probability of unanimity on the Court.[6] While there are a host of potential classifiers to use in this setting, we use random forests (Breiman, 2001), a supervised classification algorithm that has enjoyed substantial recent interest in political science applications (e.g., Grimmer and Stewart, 2013; D'Orazio et al., 2014).

[6] The underlying intuition here is similar to applications of *Wordfish* (Slapin and Proksch, 2008) and *Wordscores* (Laver, Benoit and Garry, 2003), where the researcher identifies documents as falling on either end of a dimension of interest in order to extract an underlying "explanatory" dimension.

To train the classifier, we acquired the texts of all Supreme Court cases from 1953 through 2000 from `public.resource.org`, an online repository of government documents. To get opinion-level data, we wrote a computer program in `Perl` 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.[7] We then matched the opinions to the extensive case information available from the Supreme Court Database (Spaeth et al., 2012). We retained only majority opinions decided unanimously with 9 votes, and dissenting opinions. In addition to standard text preprocessing[8] we retained case citations, and wrote a program to identify and mark negated terms consistent with prior research (Das and Chen, 2007; Pang, Lee and Vaithyanathan, 2002). Finally, we retained as features only the top 5,000 words as determined by term frequency - inverse document frequency.

Overall, the trained classifier correctly predicts 74% of the cases. More important for our purposes are the underlying class probabilities; the probability of a case being a 9-0 majority opinion serves as an opinion-level measure of polarity insofar as dissenting opinions and unanimous majority opinions are strong signals of sentiment on the Court. To assess the validity of this approach, we then used the trained classifier to measure the polarity (probability of being a unanimous majority opinion) of all Supreme Court majority opinions from 1953 through 2000, then compare the polarity measure with previously utilized measures of comity on the Court. One such measure, as detailed above, is the number of votes in the majority coalition. In the upper right panel of Figure 2 we plot the average of standardized[9] polarity scores generated by the random forests classifier

[7] Note that our "special" category includes "in part" opinions.

[8] Specifically, we removed capitalization and most punctuation, and also stemmed the terms.

[9] These are standardized solely for this plot in order to enhance comparability with the dictionary-based method we present below.

against the number of majority votes. This initial analysis provides evidence of convergent validity, with larger opinion coalitions deemed more positive, on average.

Note that this result is not unexpected given the dimension used to train the classifier; machine learning classifiers are (unsurprisingly) good at achieving the result on which it was trained. However, if the classifier outcome does not accurately reflect the underlying concept to be measured, biases can be introduced into the analyses. Further examination reveals exactly that problem on these data. Because the classifier does a particularly good job of identifying majority opinions, it actually obscures significant variation across smaller vote coalitions. This is particularly evident in the upper left panel of Figure 2. Moreover, the most informative terms identified by the classifier, presented in Figure 3, have little relation to emotional content, but rather relate to terms disproportionately likely to appear in unanimous opinions. For instance, "improvidently" is frequently used in concert with the Court dismissing a case which it had previously agreed to hear, as when a case is "dismissed as improvidently granted."

Importantly, this is not to say that supervised classifiers are not useful; on the contrary, we believe our results show how exceptionally useful they can be. However, their accuracy demands that they be deployed with appropriate, considered, and careful attention as to the dimension of interest. Here, estimating a classifier on an outcome related, but not identical, to the dimension of interest yields a measure that, while perfectly useful if we were interested in the number of votes as a measure of polarity, are not demonstrably related to intracourt social dynamics.

Our Approach

Rather than building from potentially problematic proxies, we instead utilize the approach we outlined above. To undertake this analysis, we acquired the texts of *all* Supreme Court cases, backdating the data collection of the machine learning approach above to the very first U.S. Supreme Court case in 1792. We again utilized the `Perl` program in order to separate each case file into an opinion level file. The data thus constitute a comprehensive population of the writings of Supreme Court justices, with nearly 35,000 unique opinions spanning more than 200 years and the tenures of 16 chief justices.

We applied our two approaches – conjunctive and co-occurrence – to the resulting bodies of text to estimate the aggregate sentiment of each opinion, and in addition generated sentiment scores for each opinion using a standard pre-constructed (LIWC) sentiment dictionary. At the individual (opinion) level, the correlations between the various measures are all positive, but not especially high; among majority opinions, for example, they range from 0.11 (for LIWC-coocurrence) to 0.34 (for LIWC-conjunction). Similar correlations are observed among the other opinion types. Importantly, however, those aggregate correlations mask significant variation over time. Figure 4 shows the Pearson correlations among the three polarity measures for each case, broken down by four Court eras (1791-1849, 1850-1899, 1900-1949, and 1950-2000). Note that the strongest intercorrelations for the earlier period are for the two corpus-based dictionaries; in contrast, the LIWC dictionary (which assigns polarity based on contemporary word usage) yields measures of opinion polarity that are only very weakly correlated with those derived from the two corpus-based dictionaries. This suggests, perhaps unsurprisingly, that our corpus-based approach provides a more inter-method reliable measure of sentiment in earlier periods, when opinion language differs significantly from the modern usage on which LIWC is based.

As before, we averaged the three sentiment measures to generate a composite measure of *polarity*, this time for each opinion decided by the Court. Turning to validation, we can compare this measure to the number of majority votes in each case. To do so, we again match majority opinion texts with the information contained in the Supreme Court Database. We plot the density of polarity by the number of majority votes in the lower left panel of Figure 2 and the average standardized polarity by the number of majority votes in the lower right panel of the same figure. Beginning with the density of polarity by opinion type, note that our dictionary-based approach does not arrive at the same concentration of "positive" opinions. Instead, we find that opinions tend to be only relatively positive; the mean level of polarity for all opinions in our data is 0.179, the median is 0.186, and only about 25 percent of all opinions fall below zero polarity (that is, are more negative than positive). Moreover, the variation is consistent with patterns of majority voting. Majorities with only five or six votes tend to be the most negative, while larger majorities tend to be more positive. The results also suggest variation by opinion type that are likely to be of substantive interest to scholars studying the justices' strategies in opinion writing (e.g. Carrubba et al., 2012; Epstein and Knight, 1998). To wit, the slight increases in positivity reflected in the density plot and the averages plot in opinions with eight majority votes could reflect strategic appeals for a unanimous Court, while the slight decrease in average negativity at five vote majorities may signal a tempering of language to ensure the majority survives the opinion writing process.

Beyond being potentially useful for future analyses of opinion bargaining dynamics like those outlined above, we can also now return to the research of Danelski and subsequent legal scholars and social scientists. To that end, we plot three-year moving averages of the polarity scores for the four types of opinions in Figure 5; in each subplot, vertical dashed lines indicate years in which a new chief justice took office. We also note a slight decline in polarity over time across all opinion types, although the trend is again more noticeable in majority opinions than in others (and is almost completely absent in *per curiam* opinions). And, mean levels of polarity vary relatively little across different opinion types, ranging from 0.149 for *per curiam* opinions through 0.177 and 0.184 for dissenting and majority opinions, respectively, to a high of 0.201 for concurring opinions.

In separate work (Rice and Zorn, 2013), we analyze the marginal effect of changes in the chief justiceship on the polarity of individual justices' opinions. We find that marginal shifts in linguistic polarity associated with changes in the chief justice are seen to be small and largely random in *per curiam* opinions, though we do note that – consistent with much previous work – the only value to fall below zero on this measure is for Chief Justice Stone. Among majority opinions, we find the highest levels of positive language during the tenures of chief justices Marshall, Stone, and Hughes, and the lowest for justices White and Warren. Consistent with much earlier work (Danelski, 1960; Walker, Epstein and Dixon, 1988; Haynie, 1992), the appearance of Justice Stone in this list may suggest something about the dynamics of disagreement on the Court during his era: only in those instances where an opinion is issued for the entire Court do we find greater negativity during the Stone era.

Summary and Future Directions

Our goal at the outset was to develop a method for building sentiment dictionaries that yield valid, reliable measures of sentiment for corpuses of specialized language, and to do so in a way that minimizes the amount of human coding necessary. Such a method would be very valuable for analyzing text where standard plain-language sentiment dictionaries fail. We characterize our approaches as "minimally supervised" (e.g., Uszkoreit, Xu and Li, 2009) in the sense that they require a small amount of initial human coding but are largely unsupervised in nature. Our work here indicates that such dictionaries do a credible job of recovering sentiment, and that they may be especially useful in circumstances where language is specialized and/or when its use changes over time.

In closing, we note that a number of additional analyses are necessary before the full value of our approach can be determined. Chief among these is a more thorough validation of the dictionaries our methods yield; this would include more detailed comparisons of conjunction-based dictionaries with those derived from co-occurrences, more detailed benchmarking against existing human-coded dictionaries for specialized vocabularies, and – where possible – additional efforts to validate sentiment scores derived from those dictionaries against other well-understood sentiment measures (either quantitative measures or those based on supervised-learning approaches). A related question is dictionary size: While our conjunction-based approach yields a "natural" size for the resulting dictionary, our co-occurrence method requires use of a stopping rule to determine the size of the dictionary. At present there is little clear guidance about the optimal dictionary size for sentiment analysis; at a minimum, then, we plan on testing the sensitivity of the various dictionaries to different stopping rules.

We also foresee 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 in different specialized contexts. If in a particular context, for example, speakers tended not to "string" adjectives and adverbs together with conjunctions, the usefulness of our conjunctive approach would be attenuated. To address this, we plan to apply our methods to a range of different corpuses from various scientific, literary, and artistic fields of endeavor, and to texts drawn from both formal (e.g., official documents) and informal (message boards, blog posts) sources. Similarly, because our initial findings provide some evidence that our approaches may have advantages when applied to texts from earlier eras, we also plan to compare the performance of dictionaries constructed using our methods to standard ones as applied to older corpuses.

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.

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

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

Carrubba, Cliff, Barry Friedman, Andrew Martin and Georg Vanberg. 2012. "Who Controls the Content of Supreme Court Opinions." *American Journal of Political Science* 56(2):400–412.

Corley, Pamela, Amy Steigerwalt and Artemus Ward. 2013. *The Puzzle Of Unanimity: Consensus On The United States Supreme Court*. Stanford University Press.

Corley, Pamela, Udi Sommer, Amy Steigerwalt and Artemus Ward. 2010. "Extreme Dissensus: Explaining Plurality Decisions on the United States Supreme Court." *Justice System Journal* 31(2):180–201.

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.*

Das, Sanjiv and Mike Chen. 2007. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science* 53(9):1375–1388.

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.

D'Orazio, Vito, Steven Landis, Glenn Palmer and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 44:forthcoming.

Epstein, Lee and Jack Knight. 1998. *The Choices Justices Make*. Washington, DC: Congressional Quarterly.

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.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 2007. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics pp. 174–181.

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.

Laver, Michael, Kenneth Benoit and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 47:215–233.

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.

Owens, Ryan and Justin Wedeking. 2011. “Justices and Legal Clarity: Analyzing the Complexity of U.S. Supreme Court Opinions.” *Law & Society Review* 45(4):1027–1061.

Owens, Ryan and Justin Wedeking. 2012. “Predicting Drift on Politically Insulated Institutions: A Study of Ideological Drift on the United States Supreme Court.” *Journal of Politics* 74(2):487–500.

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.

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 and Christopher Zorn. 2013. The Evolution of Consensus in the U.S. Supreme Court. In *Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois.*

Slapin, Jonathan and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

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.

Toutanova, Kristina, Dan Klein, Christopher Manning and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*. pp. 252–259.

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.

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

Model	**Mean**	**Min**	**Max**
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
Our Polarity Approach	72.5	-	-

NOTE: Estimates for naive Bayes, maximum entropy, and support vector machines classifiers are taken from Pang, Lee and Vaithyanathan (2002).

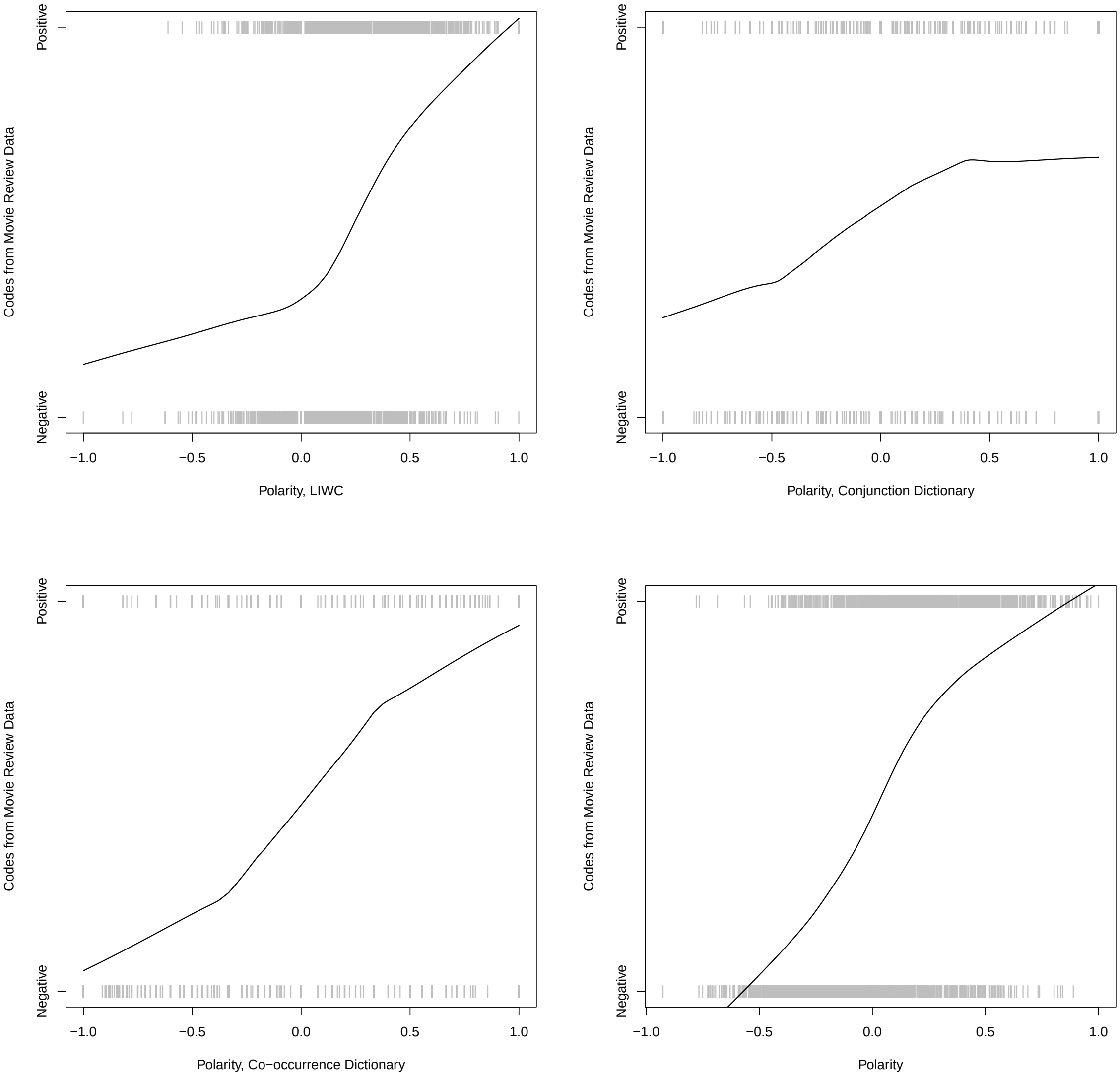

Figure 1: *Measures of polarity by assigned movie ratings (positive or negative).* Plots are estimated polarity (x-axis) by positive and negative ratings, as determined by authors. See text for details.

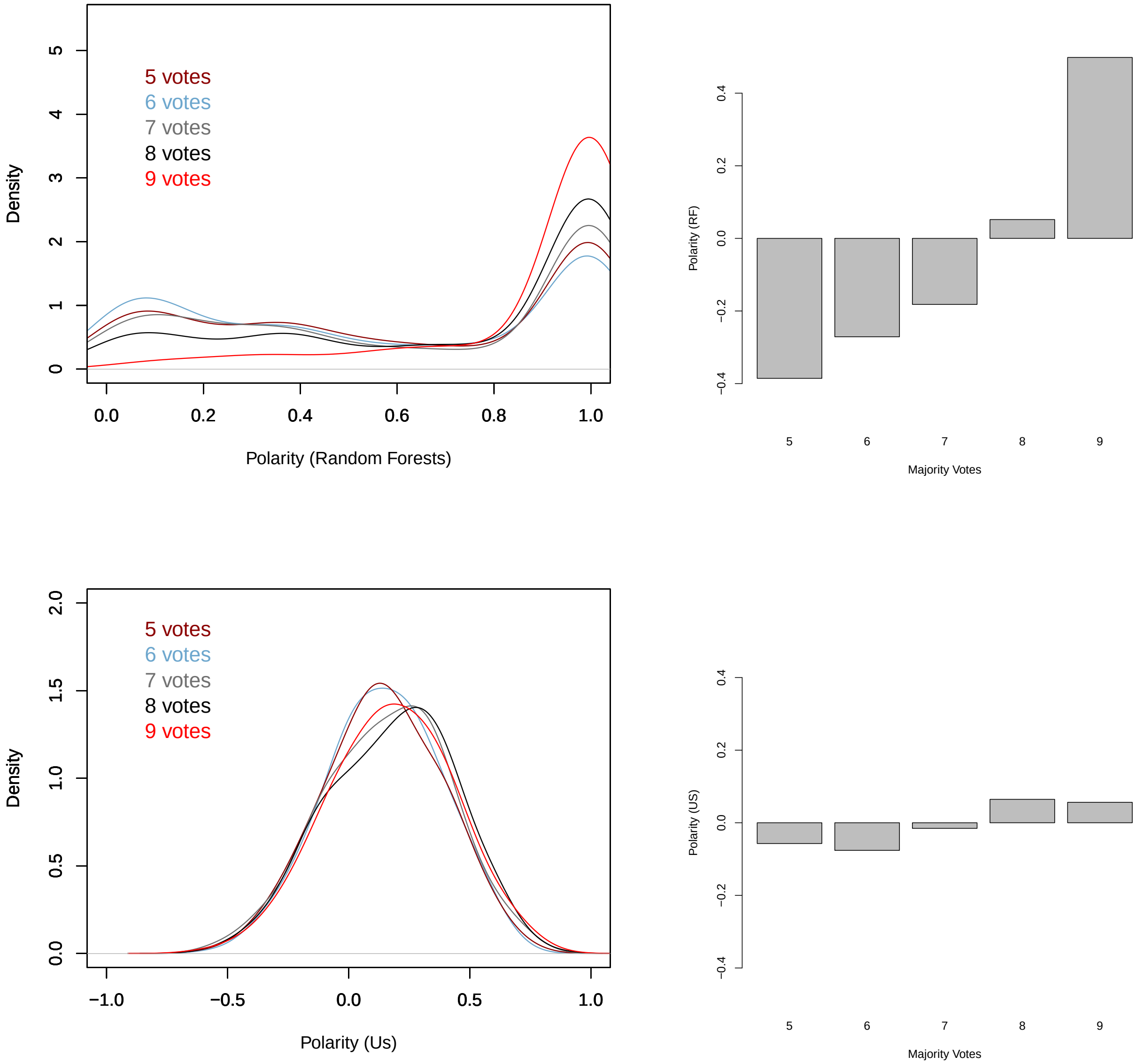

Figure 2: *Comparison of Random Forests Classifier and Our Method.*

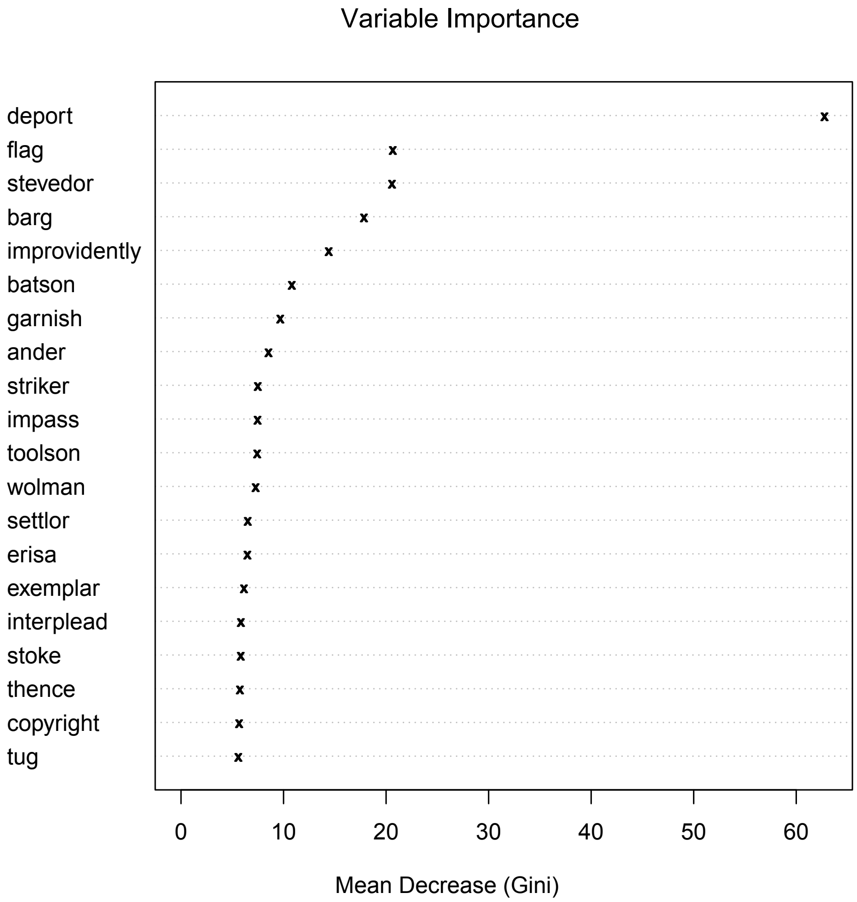

Figure 3: *Plot of Variable Importance According to Random Forests Classifier*

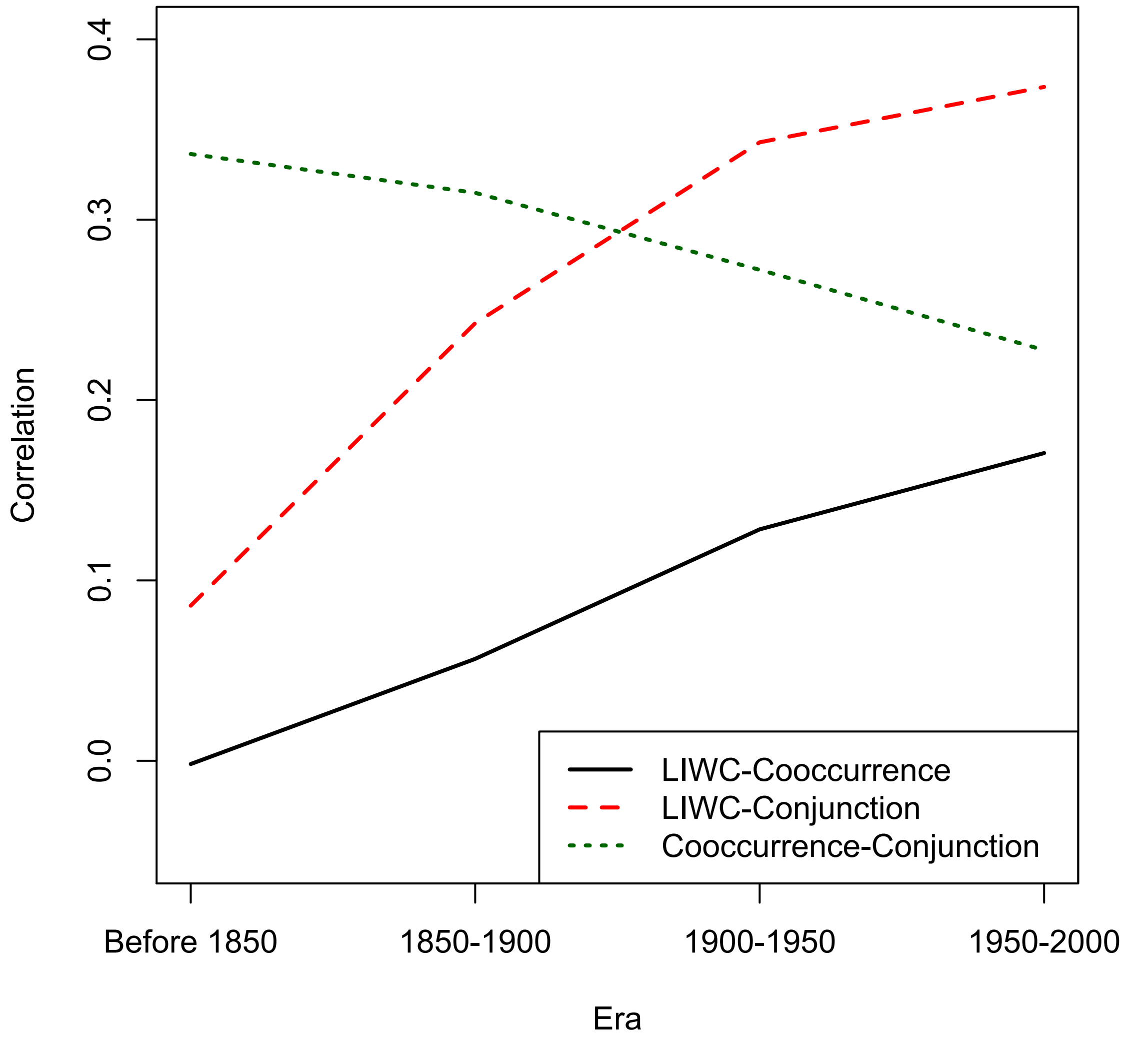

Figure 4: *Correlations among polarity measures, by dictionary type and era.*

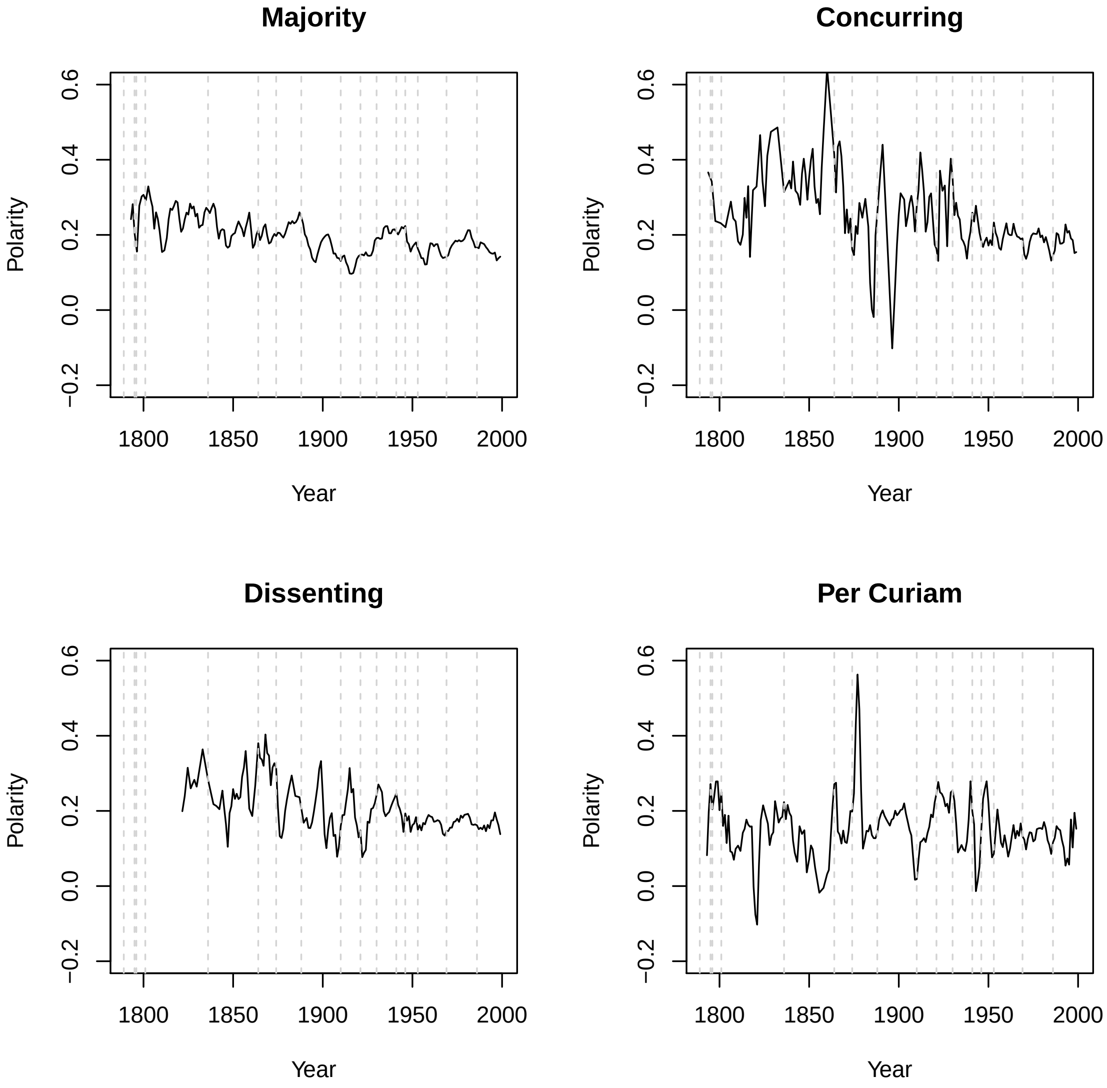

Figure 5: *Three-year moving average of opinion polarity, by year and opinion type.* Dashed vertical grey lines represent the first year of a new chief justice tenure.