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

Douglas R. Rice
Department of Political Science
University of Mississippi
douglas.r.rice@gmail.com

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

Version 0.1

September 19, 2013

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, and show its usefulness in an application to the specialized language used in U.S. federal appellate court decisions.

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

[*] Prepared for presentation at the New Directions in Analyzing Text as Data Workshop, September 27-28, 2013, London, England. This paper is a very preliminary draft; the content is likely to change with future revisions. Please be aware of this if you choose to cite this version.

(positive or negative); but it can also 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 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 or difficult to obtain. In the latter instance, acquisition of training data often requires either the subjective human-coding of a substantial number of texts or reliance on potentially inaccurate proxies of sentiment. In either case, machine learning approaches suffer from 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). While pre-constructed dictionaries offer ease of use and 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 methods. The third section describes an initial effort to validate our methods through its application to a well-understood corpus of data on film reviews. The fourth section applies our approach to an initial 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. 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 or negative irrespective of their context. We then adopt two separate techniques – word *conjunctions* 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. Another program in Perl then identified these seed words and whether they are used in conjunctions in any of the texts. All words used in conjunctions with our one

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

of the words in our seed sets are then retained for inclusion in the respective (positive or negative) dictionary. The resulting dictionaries are then cleaned of irrelevant and inappropriate terms, yielding a conjoined dictionary of context-specific terms derived directly from the texts.

In the *co-occurrence* approach, we identify words which co-occur in texts with those seed words. From those, 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. Depending on the size of the corpus and the total number of unique terms within the corpus, we then 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 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 requires 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 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 average 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] The data, introduced in (Pang and Lee, 2004), consist of 2000 movie reviews – 1000 positive and 1000 negative – pulled 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 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

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

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

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. Therefore, we 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 suboptimal, 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.

This is not to argue that our approach is a substitute for supervised machine learning approaches. 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 and available rating is available for training a classifier, our approach generates estimates close to the best-performing machine learning classifiers for this benchmark dataset. Therefore, having documented the validity of our approach, we turn next to one such research area where no natural rating of sentiment is available: 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); Black and Spriggs (2008)), but the existence of consensual norms make it likely that such indicators will mask the true level of dissensus on the Court.

Our approach relies instead on the text of the opinions – majority, concurring, dissenting, and *per curiam* – themselves, rather than on the votes of the justices. The language of opinions is the central mechanism by which the justices convey the substance of their rulings to the legal community and the public, and those opinions often contain language that – often strongly – conveys their emotional attitudes toward the decisions at hand.

To undertake this analysis, we acquired the texts of all Supreme Court cases from 1792 through 2000 through the website `public.resource.org`, an online repository

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

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. Note that the "special" category includes "in part" opinions. 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 2 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 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.

We then averaged the three sentiment measures to generate a composite measure of *polarity* for each opinion decided by the Court. Three-year moving averages of these scores for the four types of opinions are presented in Figure 3; in each subplot, vertical dashed lines indicate years in which a new chief justice took office. Note first that, on balance, opinions tend to be 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). 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, as with subjectivity, 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 "semi-supervised," in the sense that they require a small amount of initial human coding but are largely unsupervised in nature. Our preliminary work here indicates that such dictionaries do a credible job of recovering sentiment, and that they may be especially useful in circumstances where language 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.

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.*

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.

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.

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.*

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.

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 classifers are taken from Pang, Lee and Vaithyanathan (2002).

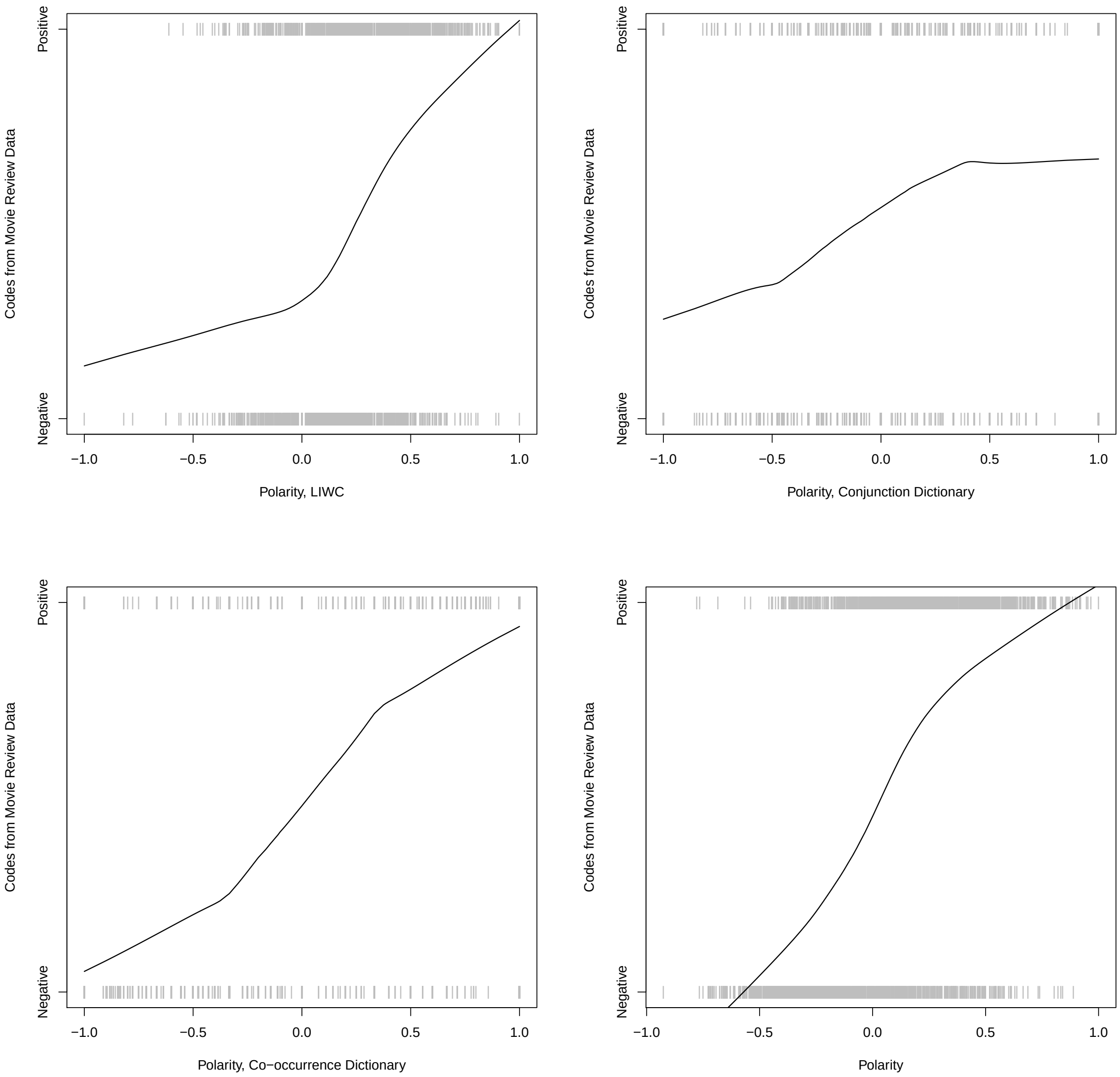

Figure 1: *Measures of polarity by assigned movie ratings (positive or negative).* Plots are estimated polarity (x-axis) by positive and negative ratings, as determined by authors. See text for details.

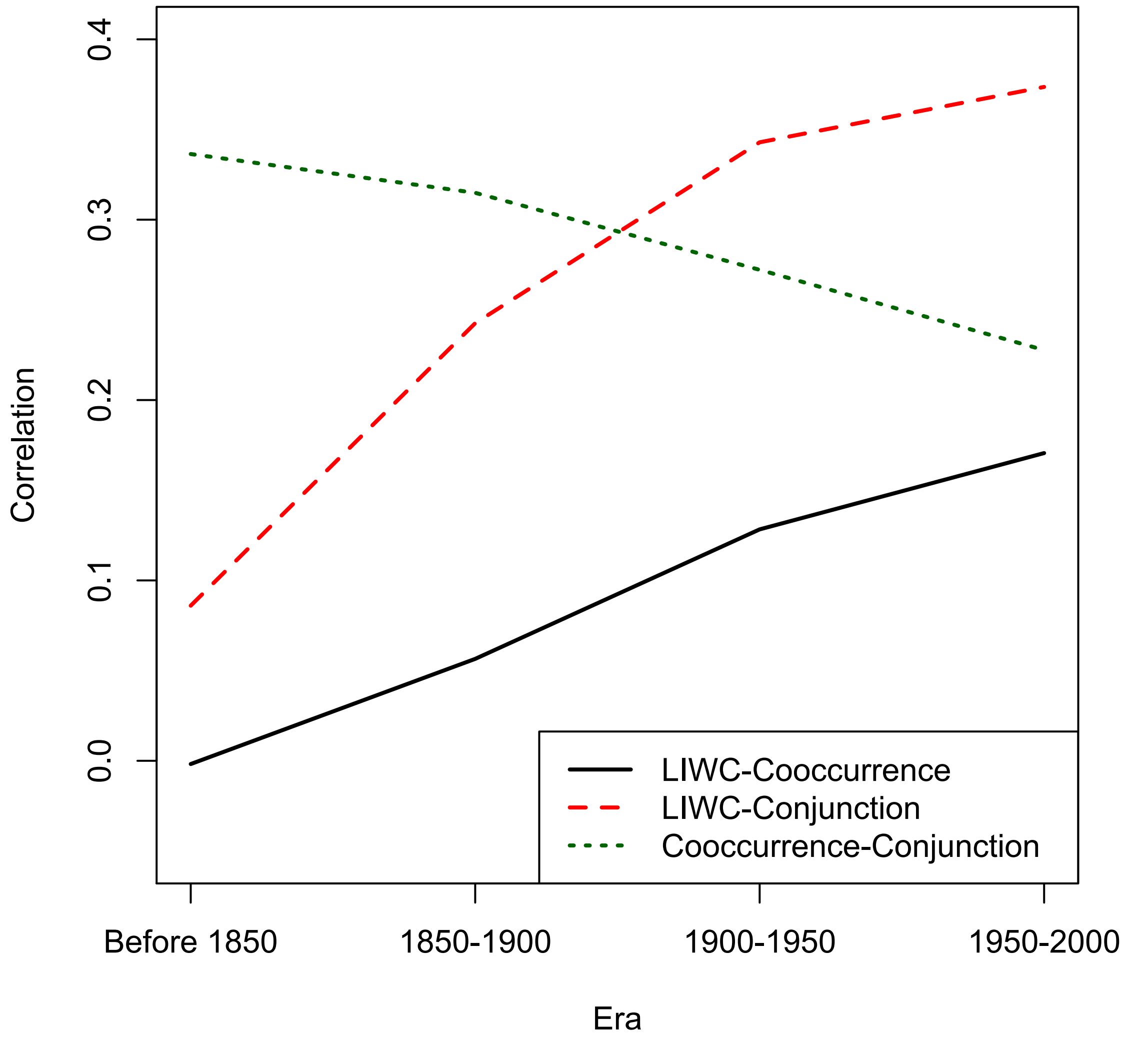

Figure 2: *Correlations among polarity measures, by dictionary type and era.*

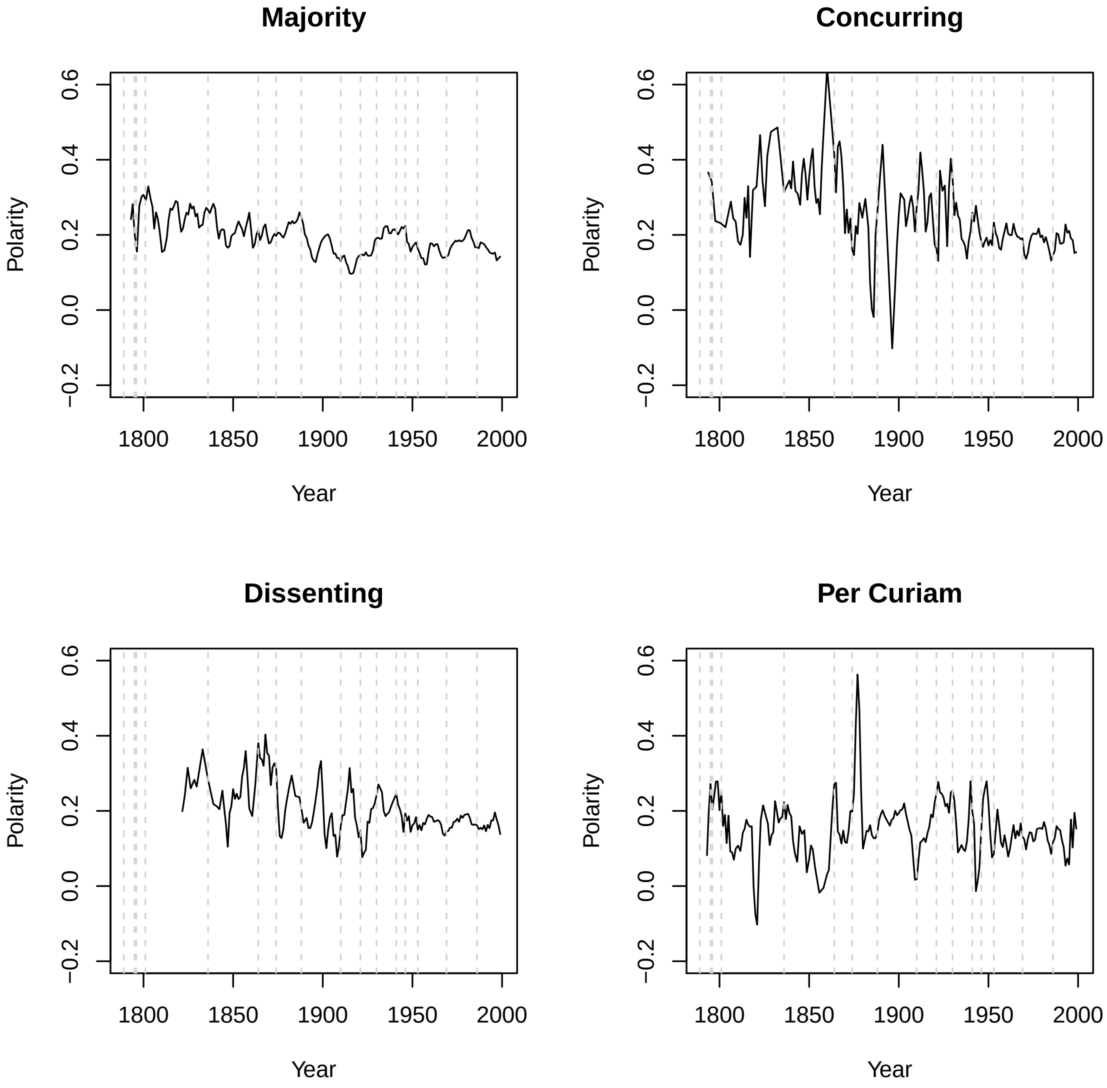

Figure 3: *Three-year moving average of opinion polarity, by year and opinion type.* Dashed vertical grey lines represent the first year of a new chief justice tenure.