Journal of Empirical Legal Studies
Volume 4, Issue 4, 1007–1039, December 2007

Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research

*Michael Evans, Wayne McIntosh, Jimmy Lin, and Cynthia Cates**

Political scientists in general and public law specialists in particular have only recently begun to exploit text classification using machine learning techniques to enable the reliable and detailed content analysis of political/legal documents on a large scale. This article provides an overview and assessment of this methodology. We describe the basics of text classification, suggest applications of the technique to enhance empirical legal research (and political science more broadly), and report results of experiments designed to test the strengths and weaknesses of alternative approaches for classifying the positions and interpreting the content of advocacy briefs submitted to the U.S. Supreme Court. We find that the Wordscores method (introduced by Laver et al. 2003), and various models using a Naïve Bayes classifier, perform well at accurately classifying the ideological direction of amicus curiae briefs submitted in the *Bakke* (1978) and *Bollinger* (2003) affirmative action cases. We also find that automated feature selection techniques can enable the detection of disparate issue conceptualizations by opposing sides in a single case, and facilitate analysis of relative linguistic "reliance" and "dominance" over time. We conclude by discussing the

*Address correspondence to Wayne McIntosh, Department of Government & Politics, University of Maryland, College Park, MD 20742; e-mail: wmcintosh@gvtpt.umd.edu. Evans is in the Department of Government & Politics, University of Maryland; McIntosh is Associate Professor, Department of Government & Politics, University of Maryland; Lin is Assistant Professor, College of Information Studies, University of Maryland; Cates is Professor, Department of Political Science, Towson University.

This research was conducted with support from the National Science Foundation, under award BSC-062467, from September 2006 to August 2009. We acknowledge a number of colleagues who have contributed in one way or another to this project. Ken Cousins, Ken Benoit, Mark Kessler, Irwin Morris, Richard Pacelle, Geoff Layman, and James Spriggs commented on earlier drafts; Rebecca Thorpe, Christina Heshmatpour, Marlaine White, and Amy Hershkovitz provided important research support. We also benefited considerably from comments and suggestions by anonymous reviewers, as well as from colleagues at the First Annual Conference on Empirical Legal Studies (University of Texas Law School, 2006).

implications of our results and pointing to areas where technical and infrastructure improvements are most needed.

I. Introduction

Students of the judicial process of every methodological inclination have at least one thing in common: they all rely on information contained in legal texts. Interpreting the legal meaning of judicial opinions is, of course, a time-honored endeavor among traditional legal scholars, but even the most quantitatively inclined empirical legal researchers rely on judicial texts in their analyses. The most notable example of such quantitative work employs Harold Spaeth's U.S. Supreme Court Judicial Database (2006), which is a trove of case variables based on the coding of every Supreme Court opinion dating back to the beginning of the Warren Court (1953).[1] Hall and Wright (2007) recently argued that the best legal scholarship combines the analytical abilities of legal experts with the scientific rigor of systematic content analysis. The former is essential for appropriate case selection and the accurate identification and coding of appropriate variables, while the latter allows researchers to test the empirical validity of conclusions drawn about relationships among those variables. Hall and Wright's examination of 166 academic legal research projects employing systematic content analysis gives credible support to their position. However, the work also demonstrates at least two perennial challenges attendant on the use of content analysis. First, projects face a tradeoff between large-scale inquiry focused on "thin" observations (e.g., voting agreement, participation, coalition size, length of legal documents),[2] and smaller-scale studies that involve the coding of more abstract and nuanced concepts. This inverse relationship between breadth and depth limits researchers' ability to understand the dynamics of the judicial system more fully. Second, content analysis almost always raises questions about coding reliability (see, e.g., Johnson 1987), and this is especially true with judicial research. Legal texts are lengthy and dense, presenting serious challenges to even the best doctrinal specialists and historians. Individual scholars find it difficult to maintain consistency when

[1] Similar databases are available for U.S. courts of appeals (Songer 1989) and state supreme courts (Brace & Hall 2000).

[2] For the purposes of this discussion, we use the terms "text" and "document" interchangeably.

coding complex documents, particularly when the objective is to compare multiple documents. This problem is even more serious for team-based approaches (e.g., Carmines & Zeller 1979).[3] We believe automated content analysis techniques hold the promise of allowing legal researchers to overcome these problems.

In recent decades, information retrieval researchers have developed techniques for efficiently storing and retrieving texts at a large scale (e.g., Frakes & Baeza-Yates 1992; Salton 1989). This is accompanied by increasingly sophisticated statistical algorithms for analyzing the structure and content of natural language text (e.g., Manning & Schütze 1999). Furthermore, computer-assisted qualitative data analysis (CAQDAS) techniques that combine automation with manual effort to take advantage of specific human-engineered domain knowledge are evolving rapidly (see Gibbs et al. 2006). For example, precoded dictionaries consisting of words and phrases theoretically linked to analytical dimensions of the coder's choosing can be automatically applied to texts (e.g., Benesh & Czarnezki 2006; Schonhardt-Bailey 2006; Coffey 2005; Schonhardt-Bailey 2005). Or, as we suggest in this article, feature selection can be used to build the dictionaries automatically, thus expanding the volume of text that can be processed. These and other computational techniques can potentially allow researchers to have the best of both worlds: reliable and detailed analyses of legal documents at a large scale.

Political scientists in general and public law specialists in particular have only recently begun to exploit machine learning techniques to assist with such research questions (e.g., Laver et al. 2003, 2006; Martin & Vanberg 2006; Benoit et al. 2005; McGuire & Vanberg 2005; Giannetti & Laver 2004; McIntosh et al. 2004). This article provides an overview and assessment of this methodology. We proceed as follows. In the next section we describe the basics of text classification. In Section III, we suggest applications of this technique to enhance empirical legal research. In Section IV, we report results of experiments designed to test the strengths and weaknesses of alternative computational models for classifying the positions and interpreting the content of briefs submitted to the U.S. Supreme Court. In Section V, we suggest broader political science applications for automated content analysis techniques and outline further work that needs to be done to maximize the potential usefulness of this methodology. Finally, in Section

[3] Also see the reliability analysis James Gibson conducted as part of the Supreme Court Database Project (Gibson 1997).

VI, we conclude with our overall assessment of these methods for enhancing social science research.

II. An Overview of Text Classification

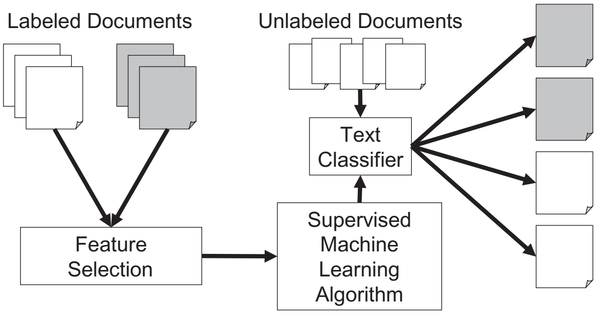
Text classification is a generic problem that lies at the intersection of computational linguistics and information retrieval (Lewis 1992; Brill & Mooney 1997; Knight 1999). It can be intuitively described as the task of automatically sorting "items" into "bins." In our domain, "items" represent legal texts and "bins" (called labels, categories, or classes) could correspond to any directly or indirectly observable characteristic, such as political ideology, issue bias, or voting behavior. The goal is to develop automatic methods for labeling previously unseen documents according to some predefined coding scheme, where the labels are drawn from a finite set of alternatives. The machine learning approach to this problem (e.g., Mitchell 1996; Sebastiani 2002) involves applying different algorithms to automatically "learn" characteristics that distinguish one type of text from another based on examples that have been labeled a priori.[4]

As Figure 1 schematically shows, the machine learning approach to text classification can be divided into two phases.

In the training phase, the system is presented with correctly labeled documents from which to learn (e.g., amicus briefs annotated by the party to the case they support). Typically, these labels are manually assigned by humans who have already analyzed the text according to some theoretically grounded classification based on a particular research question. In some cases, this analysis has already been implicitly performed as the result of some other activity, and may be stored as meta-data attached to the documents. A simple example is news stories that have already been classified by an editor; these labeled news articles can then be employed to train a classifier for the same task.

[4] Formally, text classification can be generalized as the task of assigning a real value to each pair $<d_j, c_i> \in D \times C$, where D is the domain of documents and $C = \{c_1, c_2, \ldots, c_n\}$ is a finite set of predefined labels or categories. We assume the existence of a function, called the target function, $\varphi: D \times C \rightarrow [0,1]$, which assigns the degree of membership of a document with respect to a category. This is a more general case than the situation shown in Figure 1 because it allows the possibility of assigning documents to multiple categories, as well as providing values to represent degrees of association. The result of the machine learning process is a function, called a classifier, $\varphi': D \times C \rightarrow [0,1]$, which approximates the target function, such that φ and φ' coincide as much as possible—that is, the functions assign the same value to each document-category pair.

Figure 1: The machine learning approach to text classification.



NOTE: This figure illustrates the basic elements of the machine learning approach to text classification. The task involves applying a supervised machine learning algorithm to training documents of known categorical values ("labeled documents") to train a text classifier that then estimates categorical values for (applies "labels" to) previously unseen documents.

Since computers cannot "understand" documents in the same way humans do, "learning" takes place at the level of features automatically extracted from training examples, by a representation function. A feature can be any quantifiable characteristic of the text, for example, the presence and frequency of a particular word. In this study, we employ the set of words contained in a document as its features (hence, we employ "feature," "term," and "word" interchangeably). However, since features are not equally important in discriminating between different labels, feature selection is often an important step in the training process, from the point of view of both accuracy and computational efficiency. To summarize, text represented in terms of a particular feature set, which can be thought of as a "digest" of its content, and the preassigned labels together serve as the input to the machine learning algorithms that will be used to train the text classifier.

In the testing phase, the trained classifier is presented with new unlabeled documents (naturally, previously unseen in the training examples), and the computer's task is to assign labels in a manner that is consistent with the training examples. For the case where there are only two possible labels (i.e., binary classification), the system's performance can be broken down in terms of a two-by-two contingency table indicating true positives, true negatives, false positives, and false negatives. Table 1 provides a guide that illustrates various evaluation metrics. For point of illustration, let us suppose that the system is tested for its ability to correctly classify the ideological positions of documents as liberal or conservative. In the contingency table, the columns contain model outputs (predictions) and the rows contain actual

Table 1: A Contingency Table that Illustrates Binary Classification Evaluation Metrics (Using Ideological Classification as Example)

		Predicted	
		Liberal	*Conservative*
Actual	Liberal	a	b
	Conservative	c	d

Accuracy = $(a + d)/(a + b + c + d)$.

Liberal precision = $a/(a + c)$.

Conservative precision = $d/(b + d)$.

Macro-average precision = *(Liberal precision + Conservative precision)/2.*

Liberal recall = $a/(a + b)$.

Conservative recall = $d/(c + d)$.

Macro-average recall = *(Liberal recall + Conservative recall)/2.*

NOTE: Before a machine learning model is applied to a set of truly "unlabeled" documents, it is important to test the model's performance with a set of "test documents," which are documents with a position known to the researcher but not the computer. This table delineates different performance measures commonly used by computational linguists for evaluating the performance of machine learning classifiers and that we use to assess our models.

ideological positions. Overall accuracy (precision) is computed as $(a + d)/(a + b + c + d)$, which measures the fraction of all predictions that are correct. Precision can be broken down into liberal and conservative components, that is, of the documents that the model predicts as liberal, how many are actually liberal (and same for conservative)? Liberal precision (LP) and conservative precision (CP) are computed as $a/(a + c)$ and $d/(b + d)$, respectively. The complement of precision is recall, which measures the fraction of liberal (or conservative) documents that are correctly identified. Liberal recall (LR) and conservative recall (CR) are computed as $a/(a + b)$ and $d/(c + d)$, respectively. Note that precision and recall values for both categories are necessary to fully quantify the performance of a particular model: it is rather easy to score high in each individual metric. A trivial model that identifies all documents as liberal (regardless of content) would have perfect liberal recall, although at the cost of conservative recall and liberal precision. A model that only assigns the liberal label once and does so correctly would have perfect liberal precision, but would probably suffer from low liberal recall. A "good" classifier must balance all these various issues.

To assist in the interpretation of these various evaluation metrics, liberal and conservative precision can be averaged to produce what is commonly known as macro-averaged precision (Lewis 1991); similarly, liberal and conservative recall can be averaged to produce macro-averaged recall. Macro averaging places equal emphasis on the categories, even if one label is more prevalent than the other (e.g., if there are more liberal documents than conservative documents in our data set overall). The macro-averaged values answer a somewhat different question than overall accuracy. Whereas the latter focus on each individual instances (i.e., How often does the model correctly classify any document?), the former focuses on entire categories (i.e., How does the model perform on identifying documents of a particular type?). Naturally, depending on the relevant research question, one or the other type of metric would be more appropriate.

Machine learning has emerged over the last decade as the dominant approach to tackling text classification problems (Sebastiani 2002). Previously, the most popular techniques were based on knowledge engineering, which required experts to develop classification rules manually based on careful consideration of the relevant topics. The CONSTRUE system is perhaps the most famous example of this approach (Hayes et al. 1990). However, manual rule construction is not only labor intensive, but also domain specific: any changes in category structure or subject area may require substantial reengineering of the rule set. By comparison, machine learning offers three distinct advantages: greater accuracy, reduced labor costs, and portability to different domains (Mitchell 1996).

A. Wordscores: The First Political Science Application of the Machine Learning Approach to Content Analysis

Although it is not explicitly designed as a text classifier, the Wordscores method, developed by Laver et al. (2003) to estimate the policy positions of European political party manifestos, is the automated content analysis technique best known among political scientists. One advantage of this method is that it generates interval-level ideological scores for texts along a set dimension without requiring the researcher to attribute meaning to words within the text. Indeed, one can conduct the analysis even if one does not speak the language contained within the texts. As Laver et al. demonstrate, this not only simplifies the process of content analysis, but also produces results that are more accurate than those generated by the leading alternative approach to party position estimation, the Comparative Manifestos Project, which relies on extensive human coding. In various applications, the Wordscores method has proven to be an effective automated text classification method (Laver et al. 2006; Benoit et al. 2005; McGuire & Vanberg 2005; Giannetti & Laver 2004; McIntosh et al. 2004).

The process begins with selection of "reference" (training) texts, written with a known position along a dimension of interest (e.g., ideology, policy issue field, etc.). The Wordscores program then generates a word frequency matrix for every word (feature) in the reference texts.[5] Based on the relative frequencies of each word in the reference texts and the values assigned to those documents, word scores are then calculated to represent the association between words and each document. For example, let us assume reference text RT_1 is assigned a value of −10 and reference text RT_2 is assigned a value of 10. Let us further suppose that word w_{20} is used 8 times out of 3,000 words in RT_1, and 150 times out of 5,000 words in RT_2. Since w_{20} makes up a much higher proportion of the words used in RT_2 than it does in RT_1, it will receive a word score much greater than zero, suggesting that the word is more indicative of the position of RT_2 along the given dimension. More precisely, the word score for w_{20} would be equal to:

$$\frac{8/3000}{(8/3000)+(150/5000)}(-10)+\frac{150/5000}{(8/3000)+(150/5000)}(10)=.08^*(-10)+ (.92^*10)=8.37.$$

Finally, text scores are computed for unread, uncharacterized "virgin" texts (the test examples), characterizing them with respect to the reference documents. The score given to each virgin text is simply the average of all word scores for all scored words within the text. In our example, all things held equal, a virgin text that includes w_{20} at a high frequency would receive a text score that places it closer to the value assigned to RT_2 than to RT_1 on the set dimension (see Laver et al. 2003:314–16). If, in this example, the set dimension is ideology, with lower scores representing greater conservatism and higher scores greater liberalism, then a higher use of word w_{20} in virgin text VT_1 than in virgin text VT_2 would, all else held equal, indicate a more liberal ideological position in VT_1 than in VT_2.

An advantage of the Wordscores method is that it allows researchers to measure the "certainty" associated with estimated virgin text scores. Although there are several ways to accomplish this, the Wordscores procedure automatically generates standard errors, which are based on a measure of the variance of each word's score around the virgin text score, weighted by the frequency of each scored word in the virgin text (see Laver et al.

[5] Wordscores treats any string that begins with a letter as a word.

2003:317–19). With these standard errors, confidence intervals can be constructed to assess the statistical significance of variation among text scores.

B. Text Classification with Naïve Bayes Classifiers

Researchers in information retrieval and computational linguistics have examined several approaches to text classification with machine learning techniques (see Sebastiani 2002). Methods differ with respect to the algorithm used and features selected to train the text classifier. In the analysis below (Section 4.1), we compare the Wordscores method to a Naïve Bayes classifier with feature selection. In this section, we discuss the similarities and differences between the two approaches.

In the text classification literature, Naïve Bayes classifiers are viewed as a popular baseline that is easy to implement and is often competitive with state-of-the-art techniques in terms of accuracy (Rennie et al. 2003). Its underlying assumptions are actually quite similar to those of the Wordscores method. Both are based on term frequencies (i.e., the number of times a word appears in a document), although Naïve Bayes has the advantage in that it places frequency calculations on the theoretical foundation of the laws of probability.

Let us suppose that, based on the observation of a particular word w_1 from a legal brief, document D_1, one had to guess whether D_1 is liberal (L) or conservative (C) on the death penalty. The probability that D_1 is liberal (opposes) on the death penalty can be derived by application of Bayes's theorem (Borel 1965):

$$P(L|w_1) = \frac{P(w_1|L)P(L)}{P(w_1)}.$$

The probability that a particular document containing the word w_1 is liberal on the death penalty is equal to the product of the probability that a liberal brief contains the word in question and the probability that a randomly chosen brief is a liberal brief, divided by the probability that w_1 appears in any brief. To classify the document, one would simply choose the class with the highest probability. We get $P(w_1|L)$ from the training examples: like the Wordscores method, classification decisions are ultimately based on how frequently a word appears in each type of document. A Naïve Bayes classifier aggregates evidence from multiple terms (features) based on the assumption that they are conditionally independent. This independence assumption gives the algorithm its "naïve" label because it is often violated in real-world texts (Lewis, 1998). For example, the phrase “electric chair” occurs much more frequently than the independent occurrence of the individual words “electric” and “chair.” The independence assumption simplifies the complexities of content analysis and has been empirically verified to work well in many classification tasks.

Besides making use of different learning algorithms, information retrieval researchers have also explored different techniques for feature selection and weighting to improve classification accuracy. Not all words are equally important in discriminating between different labels, and feature selection methods attempt to determine automatically the set of most useful features. Such processing is potentially important in two ways: throwing away unimportant features may reduce the noise within a data set, leading to greater classification accuracy. Furthermore, working with a smaller feature set reduces computational complexity, which may be a concern for particular types of machine learning algorithms, especially on large data sets. Feature weighting methods represent different ways for assigning values to each feature with respect to a particular document, the simplest being 1 if the term appears in the document and 0 if the term does not appear in the document (a binary weighting scheme).

As a point of comparison, the Wordscores method does not employ any feature selection at all, since it processes every word found in the reference texts. As Laver et al. (2003) readily admit, this means many of the words processed are not discriminative. For example, we have found that in affirmative action cases, both liberal and conservative groups use the word “students” with similar frequencies, so the presence of that particular word is not a good indication of a document’s ideological position. The same is true for several other descriptive words, as well as function words like “and” and “the.” Instead of specially processing such terms or excluding them all together, the Wordscores method effectively treats them as ideologically moderate by assigning word scores that approach the midpoint of the set reference text score interval. The result is that “raw” virgin text scores tend to cluster tightly toward the moderate position. Although these raw scores reflect meaningfully different positions among virgin texts, their tight clustering renders them more difficult to interpret. To remedy this, the Wordscores program outputs “transformed scores,” calculated by rescaling the raw scores according to the standard deviation of the virgin text scores. Because these are problematically dependent on the virgin texts the researcher chooses to examine, Martin and Vanberg (2006) have developed a more robust transformation procedure based on raw text scores generated for the original reference texts. Although their procedure allows for a more reliable virgin text transformation, it has the disadvantage of requiring the analyst to train on only two reference texts.

By contrast, information retrieval researchers have developed a number of empirically validated techniques for processing large quantities of text documents. One standard practice is to "preprocess" texts to remove function words and other words that are not expected to form a reliable basis for discrimination. Such terms are usually referred to as "stopwords" in the information retrieval parlance. Features are weighted by frequency of terms within documents and across the entire text collection. All things being equal, it is generally true that term frequency correlates with importance; the more often a word occurs, the more likely the document is "about" the concept evoked by that word. On the other hand, however, words that appear in too many documents are not useful for capturing textual content, since they are not sufficiently discriminative. These insights can be captured using *tf.idf* term weighting, which is commonly used in many information retrieval tasks (Salton 1975; Robertson 2004). With this method, a feature (i.e., word) is assigned a weight equal to the product of its term frequency (*tf*) and inverse document frequency (*idf*):

$$\begin{aligned} w_{i,j} &= tf_{i,j} \times idf_i \\ tf_{i,j} &= c_{i,j} \\ idf_i &= \log(N/d_i), \end{aligned}$$

where $c_{i,j}$ = number of occurrences of term i in document j

N = number of documents in the collection

d_i = number of documents where term i occurs.

Beyond discarding stopwords, various feature selection techniques can be employed to choose the set of terms that are most discriminative with respect to the class labels at hand. In the binary classification case, the intuition is relatively simple: words commonly used by one side, but not the other, represent the best features with which to represent documents. Several statistical measures quantify this basic idea, for example, information gain, chi^2, odds ratio, just to name a few (Manning & Schütze 1999). The classic work of Yang and Pedersen (1997) examines different feature selection techniques and finds information gain and chi^2 to be the most effective; see also more recent work by Forman (2003). In this work, we employ the chi^2 method, which can be computed directly from a two-by-two contingency table of a term and a label.

The techniques discussed in this section can be applied to any collection of documents a researcher may wish to explore. We have selected legal texts for experimentation and turn our attention to the application of these methods to empirical legal research in the following sections.

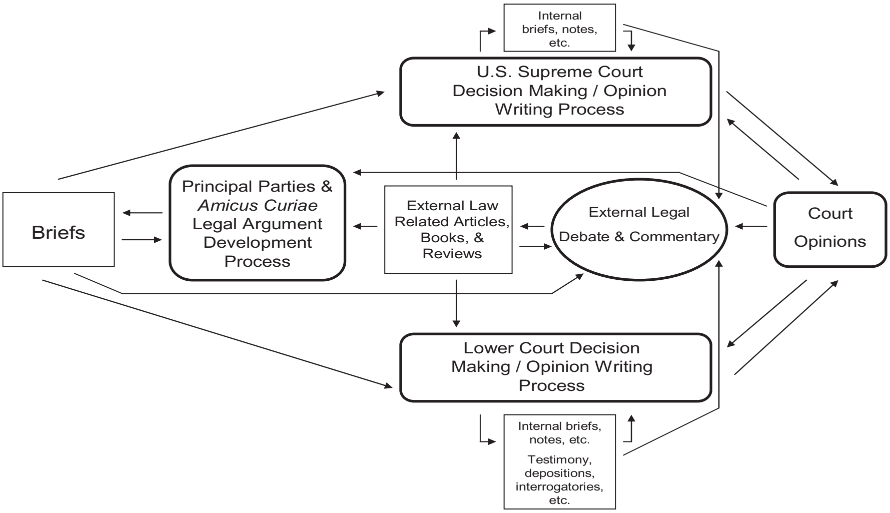
III. How Machine Learning Can Enhance Empirical Legal Research

The U.S. legal system involves intricate communication, processing, and transfer of information. It consists of agents (e.g., lawyers, judges, litigants, interest groups) whose behavior is affected by a range of influences (e.g., political ideologies, historical precedent, current political and economic context, institutional structure). The law, expressed in judicial opinions, constantly evolves to address emerging challenges and opportunities, leaving textual records that explain who won and why. In turn, these records are referenced by legal agents in the future. As such, the judicial process is a rhetorical one, expressed through text, and organized in a hierarchical institutional structure where past decisions inform present conflicts.

Figure 2 depicts the U.S. legal process as an autonomous system of textual inputs and outputs. The depiction is "autonomous" in the sense that the system is portrayed as independent of the larger political, economic, and social context in which it operates. We adopt this view for the sake of parsimony.[6] As Hall and Wright have argued, "[t]he content of judicial opinions can be important in the study of the broader social, economic, and political systems with which judicial law interacts, but cases are also well worth scientific study in their own right" (2007:32). Legal texts—in the form of principal party and amicus curiae briefs, "external" law-related publications, and court opinions—are conceptualized as both inputs into, and outputs from, text generation processes. Although a time dimension is not provided in this depiction, the schematic should be interpreted by imagining textual inputs (i.e., outputs from time t_{n-1}) flowing through the system over

[6] We, of course, do not deny the importance of extra-legal influences on the legal process or suggest that texts do not contain information about these influences also warranting systematic assessment.

Figure 2: The legal process as an autonomous system of textual inputs and outputs (with no time dimension).


Internal briefs, notes, etc.
U.S. Supreme Court Decision Making / Opinion Writing Process
Briefs
Principal Parties & *Amicus Curiae* Legal Argument Development Process
External Law Related Articles, Books, & Reviews
External Legal Debate & Commentary
Court Opinions
Lower Court Decision Making / Opinion Writing Process
Internal briefs, notes, etc.
Testimony, depositions, interrogatories, etc.

NOTE: Legal texts are here conceptualized as both inputs into, and outputs from, text generation processes. The schematic should be interpreted by imagining textual inputs (i.e., outputs from time t_{n-1}) flowing through the system over time. Each arrow pointing away from a text generation process to a document type represents the causal influence of the process and its relevant textual inputs. Similarly, each arrow pointing from a document type to a text generation process represents potential influence on the process and, by extension, its textual output.

time. Each arrow pointing away from a text generation process to a document type represents the causal influence of the process and its relevant textual inputs. For example, a legal scholar will formulate an innovative legal argument (process) in a law review article (output) after reading previous law reviews and judicial opinions (inputs). Similarly, each arrow pointing from a document type to a text generation process represents potential influence on the process and, by extension, its textual output. So, for example, an innovative law review article (input) might influence the argument used by a litigant and/or third party (process) in a brief submitted for a case (output). Each depicted causal relationship between processes and document types—including extended causal chains—entails multiple potential research agendas that could be enhanced, or permitted for the first time, by using automated text analysis techniques. Probably the most commonly studied relationship, for example, is that between judicial decision making in a case and the content of opinions from previous cases. Greater ability to make use of the information in judicial opinions can only enhance this time-honored research tradition. In general, as we develop the capacity to automatically classify and extract pertinent information from the content of legal documents accurately and reliably, we will become better able to observe, and eventually explain, policy change throughout the legal system.

Here, we briefly discuss two ways that we think interesting aspects of the legal system can be better analyzed through the application of automated content analysis techniques.

A. Legal Text Classification with Machine Learning: Measuring the "Inputs" and "Outputs" of the Legal System

For practical reasons, empirical legal researchers have not yet been able to leverage the inferential advantages of large *N* statistical analysis to build and test models of nonbinary ideological positions articulated through legal briefs and opinions. Although many scholars have used a variety of methods, including statistical analysis of content codes, to examine the *process* by which justices craft opinions (e.g., Murphy 1964; Howard 1968; Rohde 1972; Rohde & Spaeth 1976; Maltzman & Wahlbeck 1996; Epstein & Knight 1998; Wahlbeck et al. 1998; Maltzman et al. 2000), no one to date we are aware of—with the exception of McIntosh et al. (2004) and McGuire and Vanberg (2005)—has ventured to assign nonbinary ideological values to legal documents in order to build and test models that explain the policies they espouse. The problem with this, as Shapiro (1968:39) has pointed out, is that "the opinions themselves, not who won or lost, are the crucial form of political behavior by the appellate courts, since it is the opinions which provide the constraining directions to the public and private decision makers who determine the 99 percent of conduct that never reaches the courts." Although courts may determine which side prevails today, the language articulated in their opinions lives on, influencing, among others, future judges, litigants, and amici, who in turn can impact future court "outputs."

If the textual inputs and outputs depicted in Figure 2 can be reliably and meaningfully quantified, then a variety of innovative research questions can be addressed. What explains the ideological positions of the briefs submitted by litigants to a case? Are they influenced by positions taken by today's median justice in his or her opinion in a previous case? How do litigants' positions compare to the positions taken by amicus curiae? Do different types of interest groups submit more or less ideologically extreme amicus curiae briefs? How do repeat players' positions vary over time? Under what conditions (e.g., case salience, coalition size, type of opinion, position/ clarity of relevant precedent) do justices articulate extreme or moderate positions? Do lower court opinions exhibit ideological shifts in response to change in Supreme Court policy? Can litigant success be explained by the positions taken in their briefs?

Political scientists are only beginning to develop computational techniques capable of facilitating such an ambitious research agenda. Preliminary work by McGuire and Vanberg (2005) demonstrates both the promise and limitations of using the Wordscores procedure to generate meaningful scores for Supreme Court opinions. Below, we test the ability for Wordscores and the Naïve Bayes approach to classify amicus curiae briefs. Our results, as well as those of McGuire and Vanberg, point to the potential for enhancing research by applying automated content analysis methods to legal documents.

B. Content Analysis Assisted by Feature Selection and Other Automated Techniques: Observing and Interpreting Lexical Variation in Legal Texts

An advantage of the machine learning approach to text classification is that it allows one to identify meaningful variation among texts by statistically analyzing patterns in the texts' usage of features (e.g., words) *without attempting to interpret the meaning of those features.* In previous sections, we discussed how empirical legal research could be served by the ability to correctly classify legal documents. Here, we consider a different application of automated content analysis techniques: starting from categorically distinct texts, we can use feature selection methods to identify distinctive terms (which are otherwise difficult to detect), serving as a first step toward interpretive analysis. Since the computer does not truly "understand" the texts, there is little certainty that discriminative terms are in any way meaningful. However, such observations serve as a valuable point of investigative departure for a trained legal scholar. Content analysis programs often include a keyword-in-context (KWIC) function, which allows terms to be viewed in their natural context within a particular document, making such exploratory inquiries more effective.

Consider, for example, the finding by McGuire and Vanberg (2005) that using *Terry v. Ohio* (1968) (less conservative) and *National Treasury Employees Union v. Von Raab* (1989) (more conservative) as reference texts generated Wordscores text scores that make good spatial sense for 17 other conservative search and seizure decisions. While this is exciting in its own right, we might also decide to probe deeper by looking at the features driving the result. What words or phrases render a search and seizure decision more conservative than another? How do these compare to words that render decisions more moderately conservative? What does this tell us about legal rhetoric, the relationship between ideas and political-legal development, and/or the psychology of judicial decision making? As we demonstrate below, a similar application can allow for the observation and analysis of sharp differences in modes of argumentation by conservative and liberal litigants and third parties in affirmative action cases.

Furthermore, with a document selection method akin to "precision matching,"[7] one could attempt to isolate many other categorical variables to discover lexical distinctions across values for those variables. For example, one could use feature selection techniques to isolate words and phrases distinctive of economic versus social liberalism by processing documents similar in all respects (e.g., all are "liberal," same venue, same type of document, in cases of similar salience) except for issue area. Similarly, one might carefully select documents so as to isolate features indicative of overturning precedent, litigant (as opposed to amicus) rhetoric, minimum winning versus more consensual opinion coalitions, publicly salient versus nonsalient cases, and so on. As we demonstrate below, strategic document selection can also be used to assess continuity and change in language usage *over time* by different sides in a single issue area, and thus allow for tracking influence and/or legal evolution in a new way.

A final application of feature selection techniques is to use them in the beginning stages of semi-automated content analysis approaches already employed by political scientists. Such approaches involve the use of predefined dictionaries or search expressions in order to take a first cut at coding a set of documents (e.g., Coffey 2005) or to select desired documents from a database, such as Lexis-Nexis. Such dictionaries and search expressions can be used to identify several theoretically important aspects of legal documents, such as the theories of interpretation (e.g., Benesh & Czarnezki 2006), "jurisprudential regimes" (Richards & Kritzer 2002), or extra-legal sources (e.g., Bernstein 1968; Acker 1993) used by judges. The terms selected for such dictionaries and search expressions are obviously important in these types of studies. By applying automated content analysis to a set of documents whose central position is known, one can discover necessary and/or sufficient search

[7] Precision matching is a research design where the effects of independent variables are tested by carefully selecting cases that are identical in all other relevant respects (see, e.g., Epstein & Rowland 1991; Songer & Sheehan 1993). The method was first developed by psychologists who sought to discern "nurture" effects by comparing characteristics of adult identical twins separated at birth.

terms. Furthermore, by comparing a sample of known “hits” with a sample of known “near-miss” documents, one can discover helpful “and not” Boolean search terms in order to reduce the number of false positives generated by a set of search expressions or dictionary coding rules.

Of course, for automated content analysis techniques to assist with these research objectives, one must demonstrate them to be an accurate and reliable instrument for classifying texts and identifying discriminative features. In the next section, we examine the performance of two different content analysis methods as applied to legal documents.

IV. A Case Study

In the following sections, we present a case study of the ideas discussed above, assessing the performance of the Wordscores and Naïve Bayes methods at analyzing U.S. Supreme Court litigant and amicus curiae briefs. Specifically, we examine the ability of the two approaches to (1) accurately classify the ideological position of the various legal briefs, (2) identify words from those briefs that are distinctive to opposing ideological positions in enhancing interpretative analysis, and (3) detect patterns in language usage over time by advocates on a single issue.

We choose to focus on legal briefs for three reasons. First, these texts are of great intrinsic interest to scholars of the judicial process but are underanalyzed, in part because we have lacked the technological capacity to process them on a large scale. Second, because amici almost always publicly declare their support for one side or another, the documents contain an objective “ground truth” by which to judge the effectiveness of different text classification methods. Third, since amici, and perhaps even litigants, are less constrained than judges by legalistic norms, and may have to consider the effect of their rhetoric on their ability to attract and retain support by their ideologically motivated membership base (Hansford 2004), amicus briefs are better suited than opinions for testing the usefulness of feature selection techniques for identifying words that are distinctive to opposing ideological positions.

The briefs used for our experiments come from two affirmative action cases: *Regents of the University of California v. Bakke* (1978)[8] and *Grutter/Gratz v.*

[8] 98 S. Ct. 2733.

Bollinger (2003).[9] Both attracted an unusually large number of amicus briefs. All told, *Bakke* included 57 amicus briefs (15 for the conservative side and 42 for liberals) and *Bollinger* received 93 (19 conservative and 74 liberal).[10] This supplies us with plenty of test data. Furthermore, the fact that the issue encompasses deep ideological and social divisions makes the content of these briefs especially fertile for interpretive analysis. Finally, since they cover the same issue area and are 25 years apart, the cases provide an opportunity to analyze linguistic change over time.

A. *Text Classification*

1. Experiment Design

How well can one automatically determine the ideological position of legal briefs using automated content analysis techniques? How does the classification accuracy of the Wordscores method compare to that of a Naïve Bayes classifier? Our experiments employ two separate data sets (from the two above-mentioned affirmative action cases). For each set, we use the principal litigant briefs as training/reference texts and the cases' amici as the test/virgin texts. We compare two Wordscores models with four Naïve Bayes models. As Table 2 summarizes, models are distinguished according to method (Wordscores or Naïve Bayes), regardless of whether confidence intervals are taken into account (only applicable for the Wordscores method), whether texts are preprocessed with an exclusion dictionary, the number of features used in the analysis, and how the terms are weighted (only applicable for Naïve Bayes).

[9] For the purposes of this analysis, we pool the two cases from the *Bollinger* "twin bill" together (*Grutter* is at 123 S. Ct. 2325 and *Gratz* is at 123 S. Ct. 2411).

[10] Although the label "liberal" or "conservative" might be controversial when applied to any given group, well-established criteria are commonly used by judicial behaviorists for characterizing the ideological direction of a justice's votes. We simply apply those criteria to these briefs in order to give a short-hand description of their declared position on affirmative action. For our documents, all of which are from affirmative action cases, it is uncontroversial to classify an amicus group as "supporting the liberal party," even if the document was (for example) written by Exxon Mobil, so long as the group declares support for a university's (e.g., the University of Michigan's) affirmative action policies. Similarly, if the Center for New Black Leadership writes an amicus brief opposing a university's affirmative action program, then we can meaningfully classify the document as "conservative," even if the group itself would not best be described as conservative.

Table 2: Text Classification Model Descriptions

Method	*Wordscores*		*Naïve Bayes*			
Model	*WS 1*	*WS 2*	*NB 1*	*NB 2*	*NB 3*	*NB 4*
Confidence intervals?	No	Yes	N/A	N/A	N/A	N/A
Exclusion dictionary?	No	No	Yes	Yes	Yes	Yes
# Features	All[b]	All[b]	200[a]	200[a]	All[b]	All[b]
Feature weighting	N/A	N/A	Binary[c]	*tf.idf*[d]	Binary[c]	*tf.idf*[d]

[a] Features selected according to chi^2 values.

[b] For Wordscores models, "All" means all strings that begin with a letter, whereas for Naïve Bayes models it means all strings not excluded by exclusion dictionary.

[c] With binary weighting, only the presence or absence of a term within a document is recorded.

[d] With *tf.idf* feature weighting, each feature is given a weight equal to the product of its term frequency (*tf*) and inverse document frequency (*idf*).

NOTE: This table describes the parameters for our six text classification models. The Wordscores-based models were calculated with confidence intervals taken into account (WS1) and without (WS2), and neither use feature exclusion, selection, or weighting. The Naïve Bayes models are distinguished according to the number of features used in the analysis (200 highest chi^2 words in NB1, NB2; all words in NB3, NB4), and how the terms are weighted (binary in NB1 and NB3; *tf.idf* weighting in NB2 and NB4).

Our experiments examine the ability of the models to perform *binary* classification: Was a given brief written in support of the "liberal" or "conservative" position on the issue? To perform binary classification with Wordscores, we assign a value of −1 to conservative reference texts and 1 to the liberal reference texts. Virgin text scores with negative values are classified "conservative" and virgin text scores greater than zero are labeled "liberal." We believe that although the Wordscores method was developed to generate quantitative estimates of the *relative* policy positions of texts, there are good reasons for comparing the performance of each method at *binary* classification. For one thing, binary classification has the advantage that since almost all amici explicitly declare their support for one side or the other, an uncontroversial "ground truth" is readily available for a large number of these legal documents.[11] Independently generating a baseline estimate of each document's *degree* of liberalism or conservatism, by contrast, would require confronting the thorny content analytical issue of coding reliability assessment. As Johnson has noted, when:

[11] See note 14.

> content-analytic data is based on the judgments of coders reviewing textual material . . . judgments of a single coder are usually viewed as unacceptable. Panels of coders are, therefore, the norm and reliability questions usually turn on the degree of agreement/disagreement among coders. (1987:175)

In addition to using multiple coders, reliably estimating baseline standards requires taking steps to assure that all coders remain independent from each other and that coding rules are developed by different sets of coders (Johnson 1987:177). Previous tests of the Wordscores method have not demonstrated such standards of reliability testing. Although we could offer our judgments of where each brief should be placed relative to the others, we find it preferable to compare classification accuracy with respect to a reliable standard. Furthermore, this heightened reliability comes at little cost. Accurate binary classification is necessary (although, quite obviously, not sufficient) for continuous estimation of ideological position. Relative success at the former is therefore indicative of performance on the latter. This is especially true in light of the fact that, as Table 3 reports, the estimates generated by each method are highly correlated. This means that they are quite similar with respect to the relative placement of texts. The most important difference, therefore, may very well lie not in how well each method identifies briefs as more or less liberal, but rather in how well they accurately place each brief to the left or right of the absolute center of the set dimension.

Table 3: Pearson's *R* Correlations Between Wordscores' Virgin Textscores (WS1 Model) and the Position Estimates of Our Four Naïve Bayes Models for Briefs from Both *Bollinger* and *Bakke* Cases[a]

		Naïve Bayes Models			
		NB 1	*NB 2*	*NB 3*	*NB 4*
Model WS1	*Bollinger*	0.78	0.81	0.82	0.78
	Bakke	0.55	0.66	0.64	0.65

[a]Litigant briefs used as training documents and amici briefs used as test documents. Scatterplots (available on request) indicate that the linear association assumption is satisfied. All correlations are significant at (at least) the 0.0001 level.

Note: This table demonstrates that the Wordscores continuous text value attributions are highly correlated with the values generated by our Naïve Bayes models. This indicates that although our Naïve Bayes models are not designed to produce interval-level position estimates, they can do so (like Wordscores) and a natural way to compare their performance with our Wordscores models is to thus look at how well each performs binary classification (since their interval-level estimates are similar).

Table 4: Text Classification Performance, Trained on Litigant Briefs and Tested on Amicus Briefs: *Bollinger* (Set 1) and *Bakke* (Set 2) Cases

Set 1: *Bollinger* Briefs

	Wordscores		*Naïve Bayes*			
	WS1	*WS2*	*NB1*	*NB2*	*NB3*	*NB4*
Accuracy	0.860	0.851	0.828	0.828	0.892	0.871
Liberal precision	1.000	1.000	0.903	0.854	0.900	0.878
Conserv. precision	0.594	0.581	0.571	0.636	0.846	0.818
Macro-Avg. Precision	0.797	0.790	0.737	0.745	0.873	0.848
Liberal recall	0.824	0.812	0.878	0.946	0.973	0.973
Conserv. recall	1.000	1.000	0.632	0.368	0.579	0.474
Macro-Avg. Recall	0.912	0.906	0.755	0.657	0.776	0.723

Set 2: *Bakke* Briefs

	Wordscores		*Naïve Bayes*			
	WS1	*WS2*	*NB1*	*NB2*	*NB3*	*NB4*
Accuracy	0.821	0.836	0.684	0.684	0.772	0.807
Liberal precision	0.778	0.795	0.722	0.722	0.872	0.943
Conserv. precision	1.000	1.000	0.000	0.000	0.556	0.591
Macro Avg. Precision	0.889	0.898	0.361	0.361	0.714	0.767
Liberal recall	1.000	1.000	0.929	0.929	0.810	0.786
Conserv. recall	0.524	0.550	0.000	0.000	0.667	0.867
Macro Avg. Recall	0.762	0.775	0.464	0.464	0.738	0.826

NOTE: Tables report performance of our six text classification models as applied to both the *Bollinger* and *Bakke* amicus curiae briefs. This demonstrates that no model is best by all performance measures when applied to different documents. In general, however, the models all perform quite well relative to baseline expectations for random guessing (i.e., 0.50). See Table 1 for explanation of performance measures and Table 2 for specifications of the six models.

2. Results

The results of our experiments are shown in Table 4. The baseline for comparison is random guessing, which is correct 50 percent of the time.[12] In

[12] Another possible baseline is a system that always guesses the dominant class. For *Bakke*, 42 of 57 (74 percent) briefs are liberal; for *Bollinger* 74 of 93 (80 percent) of briefs are liberal; therefore, a classifier that always outputs "liberal" would achieve 79 percent or 77 percent accuracy (respectively). However, we reject this as a "fair" baseline because the a priori ideological distribution of amici briefs varies from case to case. That is, for any randomly selected case, it is difficult to predict in advance how much interest the case will attract from advocates of both ideological positions.

terms of accuracy, the best performing Wordscores and Naïve Bayes models are comparable: for the *Bollinger* set, NB3 and NB4 achieve slightly better results than WS1 (89.2 percent and 87.1 percent vs. 85.1 percent); for the *Bakke* set, both WS2 and WS1 slightly outperform NB4, the best Naïve Bayes model (83.6 percent and 82.1 percent vs. 80.7 percent). These results present evidence for the ability of automated content analysis techniques to classify the ideological positions of legal texts and point to the utility of computational techniques in general.[13]

A more complete picture emerges when we look at precision and recall broken down by label. It appears that no single model is unequivocally the "best," as each represents slightly different tradeoffs between precision and recall. From one perspective, this finding presents a problem for those who wish to apply automated content analysis techniques to analyze judicial, or any other political, documents. These results provide no guidance for the researcher who simply wishes to use the best available method to answer pertinent research questions. Although the lack of a single technique that is clearly superior to the others may pose a problem for researchers, it is far from insurmountable. First, it must be considered that all the models perform quite well in terms of overall accuracy. Second, as already mentioned, we are only in the beginning stages of exploring the space of computational techniques and we believe that over time a consensus on "best practices" will emerge. Regardless of any eventual limitations on the accuracy of classification, it should be noted that automated content analysis techniques can aid in the interpretative analysis of legal texts, a significant contribution in and of itself.

B. Feature Selection and Analysis

Next, we assess the ability of automated content analysis techniques to facilitate interpretive analysis. Feature selection methods can enable the detection and observation of distinctive lexical usage among different sets of

[13] A precursory look at the documents misclassified by three or more of our models (there were 11 such documents in the *Bollinger* set and 15 in *Bakke*) suggests that our models may have done most poorly at identifying both ideologically "extreme" advocacy groups (e.g., the Pacific Legal Foundation and NAACP) as well as (formally) ideologically "neutral" groups, such as representatives of state and federal governmental institutions (e.g., U.S. Solicitor General; governors from the States of Florida and Michigan; etc.). This most likely had less to do with our model specifications and more with our decision to use litigant briefs as training/reference documents. Future work in the application of machine learning techniques to judicial documents could benefit from closer examination of such classification errors.

documents. In our first experiment, we apply these techniques to liberal and conservative (litigant and amicus) briefs from the *Bollinger* cases in order to identify and interpret the words that are most distinctive to either ideological position. In our second experiment, so as to detect lexical change over time by opposing groups, we explore similarities and differences in word usage by affirmative action liberal and conservative parties in the *Bakke* (1978) and *Bollinger* (2003) cases.

The Wordscores method was not designed for this type of analysis, since it does not allow for easy identification of highly discriminative terms. The words with the highest word scores, it turns out, are not necessarily the most discriminative.[14] For the following analyses, we use the automated feature selection functionality in Provalis's *Wordstat* (v. 5.1.3) content analysis program. Besides easily enabling the identification of highly discriminating terms (as determined by chi^2 values), it also conveniently facilitates viewing these terms in context with its KWIC functionality and allows for categorical frequency analysis using coding dictionaries. For both experiments below, we begin by selecting the 200 words with the highest chi^2 values for all liberal and conservative litigant and amicus briefs presented in a given case.

1. Different Rhetorical Styles/Tone/Emphasis by Affirmative Action Conservatives and Liberals

Using *Wordstat*'s KWIC function, we were able to read a sampling of the most discriminative liberal and conservative words in context. A striking pattern emerged.[15] In general, liberal groups use language emphasizing the *impact* of affirmative action polices, while conservative words indicate concern over legal-constitutional limits on administrative *procedure*. High chi^2 liberal words are associated with concern about the concrete consequences of affirmative action policies on the (domestic and "global") "market" economy (and "business" interests), the "recruitment" and "training" of next-generation

[14] To take an extreme example, if a term were used once in one conservative brief, and not at all in any other (liberal or conservative) brief, it would receive a perfect conservative score, –1.0 by the values used above, despite its low overall frequency. On the other hand, a term used 510 times, but 50 times more frequently in conservative briefs (~500) than in liberal briefs (~10), would receive the slightly more liberal word score of –0.96. It seems that the second term should properly be scored as "more conservative" than the first.

[15] We want to be clear that we report our interpretations only to suggest the type of study automated feature selection can enable. To report these interpretations as "objective findings" would require independent corroboration.

"leaders" in the "labor" force (and military), and the achievement of substantial "opportunities" for all citizens, including the "poor," racial minorities, historically oppressed groups, and those from "underdeveloped"—especially "urban"—areas. Furthermore, an equality of "opportunity" and remedial, rather than strict egalitarian, conception of justice seems to inform liberals' arguments about the impact of affirmative action policies. Liberals' consequentialist orientation also incorporates arguments about the ability (and implied right) of institutions (states, acting through universities) to "shape" outcomes. The focus, nonetheless, is on the "social," economic, and "national" consequences of affirmative action policies and not on the legal-constitutional limits on the administrative procedures used to create and implement those policies. Samplings of "liberal" and "conservative" words are presented in Table 5.

In contrast, high chi^2 conservative words reflect an abstract focus on legal-constitutional justifications of, and limits on, administrative procedure; the epistemological status of social science research; and individualistic conceptions of justice. The proceduralist words take many forms, but they all are used in the context of arguments claiming that affirmative action procedures are somehow illegitimate. Some argue that the policies "unjustifiably" show "preferential" treatment toward "beneficiaries" based on "vague," "indefinite," "unreliable," and/or "amorphous" "classifications" such as "skin" color. The procedures are "forbidden" and should be "rejected," many claim, because they "violate" equal "protection" as guaranteed by the Constitution. To the extent that these words are associated with the consequences of affirmative action, the relation is based either on skepticism of liberals' claims about the "benign" impact of the policies or assertions of the perverse or "dangerous" unintended consequences of the policies.[16] Many conservatives doubt that diversity is "narrowly tailored" to achieve a compelling state interest as "purported," or that it actually delivers many of its "alleged" benefits. In fact, some argue that it unduly "burdens" the "innocent" while actually "stigmatizing" its "supposed" "beneficiaries." Another common claim is that since the history of past "discrimination" was unjust, it is "dangerous" to allow racial "categories" of any kind today. Although these and related words are combined in various ways to make several distinct arguments, a common thread uniting conservatives appears to be a heavy reliance on words that connote proceduralism, legalism, skepticism, and individualism.

[16] Indeed, many of the conservative arguments correspond with Albert Hirschman (1991).

Table 5: Sample of Words Most Discriminative of the Conservative and Liberal Positions on Affirmative Action Among Amicus Curiae and Principal Litigants in the *Bollinger* Cases

Term[a]	*Avg. Freq. per Lib. Brief*	*Avg. Freq per Cons. Brief*	*Chi*^{2}	*Interpretive Code Examples*[b]
Conservative Words				
PREFER*	2.83	41.79	39.18	Proceduralist; Race/Gender Neutral Justice
BENIGN	0.07	1.17	36.14	Intent vs. Consequences; Constraint
DISCRIM*	14.86	25.04	24.13	Proceduralist; Race/Gender Neutral Justice
PURPORT*	0.44	1.88	24.13	Skepticism
CLASSIF*	2.1	11.54	22.39	Proceduralist; Race/Gender Neutral Justice
NARROW-TAILORING	0.05	0.96	19.73	Proceduralist; Strict Scrutiny
REJECT*	2.75	7.79	19.15	Oppositional Posture
JUSTIF*	2.39	12.79	18.91	Proceduralist; Constraint
FORBID*	0.38	1.63	18.91	Proceduralist; Constraint; Race/Gender Neutral Justice
PROHIBITS	0.13	0.71	18.08	Proceduralist; Constraint
RATIONALE	0.66	5.92	17.58	Proceduralist; Legalistic
AMORPHOUS	0.25	1.29	14.62	Proceduralist; Skepticism
RACE-BASED	1.08	10.46	10.59	Proceduralist; Pejorative counterpart to liberal RACE-CONSCIOUS
Liberal Words				
LEADERS	2.70	0.13	31.03	Impact; Development
WORLD	3.00	0.42	18.74	Impact; Global
NATION*	21.0	7.04	17.90	Impact; Communitarian
IMPACT*	4.13	1.04	17.49	Impact
EFFECTIVE	2.78	0.75	16.54	Impact; Effectiveness
SOCIAL	6.84	1.71	16.05	Impact; Communitarian
COMMUNIT*	8.75	1.75	15.35	Impact; Communitarian
BUSINESS*	4.56	0.58	10.28	Impact; Efficiency; Distributive Justice
DESEGREGATION	2.34	0.17	10.24	Remedial Justice
GROW*	2.38	0.33	10.24	Change; Development
WORKFORCE	1.64	0.00	9.81	Impact; Distributive Justice; Development
RACE-CONSCIOUS	7.14	1.50	7.80	Proceduralist; Euphemistic counterpart to conservative RACE-BASED

[a] For the sake of parsimony, asterisks are used to denote lemmatized terms where morphologically-related variants are all highly discriminative. For example, "preference," "preferences," and "preferred" all had high chi^2 values and so we lemmatized them with "prefer.*"

[b] These interpretations are presented only as examples to suggest the type of study automated feature selection can enable.

Note: Highly discriminative liberal vs. conservative words (as measured by chi^2 values) identified using Provalis *Wordstat* (v. 5.1.3) content analysis program. Interpretative codes were assigned by reviewing terms in context.

This application of automated content analysis techniques could potentially free researchers from the tedium (and error proneness) of identifying the most distinguishing features of texts, thus allowing them to focus on interpreting and perhaps explaining the usage of those features in context. Although, after many hours of reading, a skilled legal scholar could undoubtedly develop a strong sense that conservative and liberal advocates did not merely disagree on the legal merits of affirmative action, but also operated from entirely different conceptualizations of the issue, it is doubtful that anyone could isolate with such precision the linguistic markers of those fundamental ideological differences without the assistance of computational techniques.

2. Detecting Differences in Language Usage by Opposing Groups Over Time

To conduct our second experiment, we first created four separate coding dictionaries, each composed of the 100 highest chi^2 liberal or conservative words from the *Bakke* or *Bollinger* cases. (For short, we refer to these as *Bakke* liberal words, *Bakke* conservative words, *Bollinger* liberal words, and *Bollinger* conservative words.) We then applied each dictionary to the *Bakke* and *Bollinger* briefs. Our primary purpose was to see which, if either, set of discriminative words in the *Bakke* case would be used more prevalently among groups in the *Bollinger* case. The intuition is that a significant adoption by one side in the *Bollinger* case of the other side's language from the *Bakke* case might indicate that the latter achieved the upper-hand on the terms of debate on affirmative action. As Table 6 indicates, *Bollinger* groups used 4.1 times as many *Bakke* liberal words as they did *Bakke* conservative words. This is preliminary evidence that *Bakke* liberals' language came to dominate the terms of the affirmative action debate in *Bollinger.* However, since at least part of this may be due to the fact that there were far more liberal briefs than conservative briefs, we also look at the proportion of *Bakke* word usage by *Bollinger* liberals and conservatives separately. Here, the evidence is still consistent with *Bakke* liberal dominance. While *Bollinger* liberals used *Bakke* liberal words at a rate of 5.3 for every *Bakke* conservative word, even *Bollinger* conservatives used *Bakke* liberal words 1.9 times more frequently than *Bakke* conservative words. However, the data also suggest that it would not be quite right to say that liberals *gained* the upper-hand over time; it seems, instead, that liberals may have deepened their advantage over time, but appear to have dominated the debate from the beginning. For example, *Bakke* conservatives used their conservative words only 1.9 times as often as

Table 6: Comparison of Relative Usage of Words Discriminative of Different Ideological Positions on Affirmative Action by Litigants and Amici in the *Bakke* (1978) and *Bollinger* (2003) Cases

	Ratio of Usage of 100 Most Discriminating Liberal Words to 100 Most Discriminating Conservative Words in . . .					
	Bakke			Bollinger		
	Liberal Briefs	*Conservative Briefs*	*All Briefs*	*Liberal Briefs*	*Conservative Briefs*	*All Briefs*
Bakke words liberal : conservative	7.3 : 1.0	1.0 : 1.9	3.6 : 1.0	5.3 : 1.0	1.9 : 1.0	4.1 : 1.0
Bollinger words liberal : conservative	4.7 : 1.0	1.5 : 1.0	3.2 : 1.0	17.7 : 1.0	1.0 : 1.6	6.4 : 1.0

NOTE: The table demonstrates discrepancies in distinctive lexical usage by liberal and conservative groups over time. The top left cell indicates that liberals in the *Bakke* case used the top 100 distinctively "liberal *Bakke* words" (i.e., words used most distinctively by liberals as opposed to conservatives in the *Bakke* case) 7.3 times more frequently than they used the top 100 distinctively "conservative *Bakke* words." The most discriminative words are identified according to their chi^2 using Provalis' *Wordstat* (v. 5.1.3). The *Bakke* briefs were input into the system in 1977–1978; they emerge again in the *Bollinger* case in 2003. The data apparently indicate dominance by the distinctively liberal words from *Bakke.*

they used *Bakke* liberal words, and *Bollinger* conservatives were even less self-reliant, using their conservative words just 1.6 times more frequently than they used *Bollinger* liberal words. By contrast, *Bakke* liberals relied on their most distinctive words more than those of *Bakke* conservatives at a ratio of 7.3 to 1, and the comparable rate for *Bollinger* liberals was 17.7 to 1. Furthermore, *Bakke* conservatives actually relied more heavily on what would later be distinctively liberal words in the *Bollinger* case than on conservative words from that case. The reliance was modest—a 1.5 *Bollinger* liberal to conservative word ratio—but nevertheless much greater than one would expect if conservatives and liberals had equal leverage on the terms of debate, especially if one considers that liberals never come close to being more reliant on (past, present, or future) conservative than liberal words. The data thus lead to the tentative conclusion that the liberal position dominated the terms of debate on affirmative action in amicus curiae and principal litigant briefs in the *Bakke* and *Bollinger* cases and that this dominance may have increased over time.

It would exceed the scope of this article to probe further, but this analysis points to several possible avenues for future legal research. One might test alternative explanations as to why conservatives use liberal language at such a high frequency while liberals barely appropriate conservative terms. For example, does this represent a conscious rhetorical strategy of cooption by conservatives? Or are conservatives simply swept away by a policy current within which they cannot avoid operating, even if they have set out to block or redirect its flow? One might also address the other side of the coin and ask why it is that liberals do not argue against conservatives on the latter's terms? Other research projects might examine relative usage by opposing sides in different issue areas. Are there areas of law where the distinctive language of the conservative position dominates? Finally, it is also possible to develop automated processes for characterizing the manner in which the same words or phrases are used by opposing sides. We could then begin to develop indices of such phenomena as "aggressive refutation" (repeating an opponent's arguments so as to refute them), "passive defense" (replying to an attack on the other's terms), "aggressive defense" (replying to an attack with counterargument on one's own terms), and so on. We then could seek to explain variation in values for the indices among different documents (briefs and judicial opinions), or use the indices themselves as independent variables in models explaining, for example, success/failure (at the merits and certiorari stage) or influence. Overall, automated content analysis techniques point to many exciting directions for future legal research.

V. Future Directions

We believe that this work points to many fruitful applications of automated content analysis, not only for empirical legal research, but also for all political scientists who engage in text analysis. Automated and semi-automated content analysis methods have already been applied to measure and explain partisan "slant" in newspapers (Gentzkow & Shapiro 2006); observe the dynamics of the U.S. Senate's agenda (Quinn et al. 2006); glean historical insights about the relative influence of interests, ideas, and institutions on Parliament's enactment of the Corn Laws (Schonhardt-Bailey 2006); measure the ideas articulated in the 2004 U.S. presidential election (Schonhardt-Bailey 2005); and measure U.S. gubernatorial ideologies (Coffey 2005). However, the potential of automated text analysis remains largely untapped. As increasingly large amounts of media material become available in digital format—and as the Internet continues to develop into a formidable political force—scholars of campaigns and the media should find much use for the computational techniques overviewed in this article. One especially exciting application would be to build on Riker's (1996) pathbreaking work on the dynamics of rhetorical strategy in political campaigns using the methods demonstrated in Section IV.B.2. These techniques may be ideally suited for measuring his concepts of "dominance" and "dispersion," thus allowing for the systematic testing and refinement of his formal theory of strategic rhetoric. In principle, this could apply to any forum where the rhetoric used by opposing positions is readily available in digital text format, including speeches, debates, media commentary, and so on. Finally, application of these tools need not be limited to the study of the art of rhetoric. For example, perhaps no one could benefit more from these analytical tools than those who search carefully for commonalities and nuanced differences among textual passages within and between great works in political philosophy.

To fully realize the potential of automated content analysis, much work remains to be done. Our experiments do not reveal a computational model that is unequivocally superior in the classification tasks we devised and there remains plenty of future research in exploring other machine learning methods for text classification. One promising avenue would be to explore the use of support vector machines, which have proven to be highly effective at classifying text in other domains (e.g., Joachims 1998). We also need to build a digital infrastructure better suited for the research interests of academics. The application of these methods for legal scholarship, for example, is currently limited by the fact that all existing text collections are designed to serve the needs of judges and lawyers. The types of large *N* content analysis projects we recommend in this article could benefit from the ability to perform search queries and from document annotations that are not currently available. For example, it is currently very time consuming to obtain (in an organized manner and useful format) all of Justice O'Connor's dissenting opinions, or all of the Pacific Legal Foundation's amicus briefs submitted to regulatory takings cases, all briefs and opinions submitted to search and seizure and death penalty cases, and so on. Finally, we need to develop a freely accessible (and preferably open-source) toolkit that will allow other researchers to easily perform the types of analyses we conduct in this article.[17]

[17] The authors are addressing many of these issues via a three-year NSF-funded initiative (BSC-0624067/September 2006 to August 2009) known as the "Digital Docket" project. See <http://www.umiacs.umd.edu/~digidock/>.

VI. Conclusion

In this article we overview the machine learning approach to text classification, offer our vision of how automated content analysis techniques can serve political science research, and report results of experiments we conducted to assess the effectiveness of different classification methods—Laver et al.'s (2003) Wordscores procedure and variants of a Naïve Bayes approach—to classify the position and facilitate interpretation of legal briefs submitted to the U.S. Supreme Court. What is the verdict as to the usefulness of automated content analysis techniques? We believe that such approaches hold great promise: not only do machine learning approaches to text classification achieve high accuracy in labeling the ideological positions of legal briefs, but feature selection methods can assist in interpretive analysis by identifying words indicative of opposing positions on a debate, allowing quantitative comparison of the articulation of distinctive words by different sides over time. This work espouses a computational approach to the analysis of political documents, whose value as a research methodology should not only be measured by the number of questions it answers but, more importantly, by the number of interesting research paths it illuminates.

References

Acker, J. R. (1993) "A Different Agenda: The Supreme Court, Empirical Research Evidence, and Capital Punishment Decisions, 1986–1989," 27 *Law & Society Rev.* 65.

Benesh, S. C., & J. J. Czarnezki (2006) "The Ideology of Legal Interpretation," presented at the Annual Meeting of the Midwest Political Science Association. Chicago, IL.

Benoit, K., M. Laver, C. Arnold, M. O. Hosli, & P. Pennings (2005) "Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Wordscoring," 6(3) *European Union Politics* 291.

Bernstein, N. N. (1968) "The Supreme Court and Secondary Source Material: 1965 Term," 57 *Georgetown Law J.* 55.

Borel, E. (1965) *Elements of the Theory of Probability*. Englewood Cliffs, NJ: Prentice Hall.

Brace, P., & M. G. Hall (2000) *State Supreme Court Project.* Available at <http://www.ruf.rice.edu/~pbrace/statecourt/>.

Brill, E., & R. Mooney (1997) "An Overview of Empirical Natural Language Processing," 18(4) *AI Magazine* 13.

Carmines, E., & R. Zeller (1979) *Reliability and Validity Assessment.* Beverly Hills, CA: Sage.

Coffey, D. (2005) "Measuring Gubernatorial Ideology: A Content Analysis of State of the State Speeches," 5(1) *State Politics & Policy Q.* 88.

Epstein, L., & J. Knight (1998) *The Choices Justices Make*. Washington, DC: CQ Press.

Epstein, L., & C. K. Rowland (1991) "Debunking the Myth of Interest Group Invincibility in the Courts," 85(1) *American Political Science Rev.* 205.

Forman, G. (2003) "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," 3 *J. of Machine Learning Research* 1289.

Frakes, W., & R. Baeza-Yates (1992) *Information Retrieval: Data Structures and Algorithms*. New York: Prentice Hall.

Gentzkow, Matthew Aaron, & Jesse M. Shapiro (2006) *What Drives Media Slant? Evidence from U.S. Daily Newspapers*. Available at <http://ssrn.com/abstract=947640>.

Giannetti, D., & M. Laver (2004) *Party Factions and Split Roll Call Voting in the Italian DS*. Working Paper, University of Bologna.

Gibbs, G. R., N. Fielding, A. Lewins, & C. Taylor (2006) *Online QDA Website*. Available at <http://onlineqda.hud.ac.uk/index.php>.

Gibson, J. L. (1997) *United States Supreme Court Judicial Database, Phase II: 1953–1993* [codebook] (Study 6987). ICPSR version. Houston, TX: University of Houston [producer], 1996. Inter-University Consortium for Political and Social Research, Ann Arbor, MI [distributor].

Hall, M. A., & R. F. Wright (2007) "Systematic Content Analysis of Judicial Opinions," *ExpressO*. Available at <http://works.bepress.com/ronald_wright/1>.

Hansford, T. G. (2004) "Lobbying Strategies, Venue Selection, and Organized Interest Involvement at the U.S. Supreme Court," 32(2) *American Politics Research* 170.

Hayes, P., P. Andersen, I. Nirenburg, & L. Schmandt (1990) "TCS: A Shell for Content-Based Text Categorization," presented at the proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications. Santa Barbara, CA.

Hirschman, A. O. (1991) *The Rhetoric of Reaction: Perversity, Futility, Jeopardy*. Cambridge, MA: Belknap Press of Harvard Univ. Press.

Howard, J. W., Jr. (1968) "On the Fluidity of Judicial Choice," 62 *American Political Science Rev.* 56.

Joachims, T. (1998) "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," presented at the proceedings of the 10th European Conference on Machine Learning. Chemnitz, Germany.

Johnson, C. A. (1987) "Content-Analytic Techniques and Judicial Research," 15(1) *American Politics Q.* 169.

Knight, K. (1999) "Mining Online Text," 42(11) *Communications of the ACM* 487.

Laver, M., K. Benoit, & John Garry (2003) "Extracting Policy Positions from Political Texts Using Words as Data," 97 *American Political Science Rev.* 311.

Laver, M., K. Benoit, & N. Sauger (2006) "Policy Competition in the 2002 French Legislative and Presidential Elections," 45(4) *European J. of Political Research* 667.

Lewis, D. (1991) "Evaluating Text Categorization," presented at the proceedings of the Speech and Natural Language Workshop. Pacific Grove, CA.

—— (1992) "Representation and Learning in Information Retrieval," Ph.D. Thesis, Department of Computer Science, University of Massachusetts.

—— (1998) "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval," presented at the proceedings of the 10th European Conference on Machine Learning. Chemnitz, Germany.

Maltzman, F., J. F. Spriggs, & P. J. Wahlbeck (2000) *Crafting Law on the Supreme Court: The Collegial Game*. Cambridge, UK/New York: Cambridge Univ. Press.

Maltzman, F., & P. J. Wahlbeck (1996) "May it Please the Chief? Opinion Assignments in the Rehnquist Court," 40(2) *American J. of Political Science* 421.

Manning, C., & H. Schütze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martin, L. W., & G. Vanberg (2006) *A Robust Transformation Procedure for Interpreting Political Texts*. Working Paper. Available at <http://polmeth.wustl.edu/retrieve.php?id=592>.

McGuire, K. T., & G. Vanberg (2005) "Mapping the Policies of the U.S. Supreme Court: Data, Opinions, and Constitutional Law," paper presented at the 2005 Annual Meeting of the American Political Science Association. Washington, DC.

McIntosh, W. V., M. C. Evans, & C. Cates (2004) "Only Words, or Data? Assessing the Relative Policy Positions in Supreme Court Briefs and Opinions," paper presented at the Annual Meeting of the American Political Science Association. Chicago, IL.

Mitchell, T. (1996) *Machine Learning*. New York: McGraw Hill.

Murphy, Walter F. (1964) *Elements of Judicial Strategy*. Chicago, IL: Univ. of Chicago Press.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, & D. R. Rater (2006) *An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th–108th U.S. Senate*. Working Paper. Available at <http://polmeth.wustl.edu/retrieve.php?id=622>.

Rennie, J. D., L. Shih, J. Teevan, & D. R. Karger (2003) "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," presented at the proceedings of the 20th International Conference on Machine Learning. Washington, DC.

Richards, M. J., & H. M. Kritzer (2002) "Jurisprudential Regimes in Supreme Court Decision Making," 96(2) *American Political Science Rev.* 305.

Riker, W. H. (1996) *The Strategy of Rhetoric: Campaigning for the American Constitution*. New Haven, CT: Yale Univ. Press.

Robertson, S. (2004) "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," 60 *J. of Documentation* 503.

Rohde, David W. (1972) "Policy Goals and Opinion Coalitions in the Supreme Court," 16 *Midwest J. of Political Science* 208.

Rohde, David W., & Harold J. Spaeth (1976). *Supreme Court Decision Making*. San Francisco, CA: W.H. Freeman.

Salton, G. (1975) "A Vector Space Model for Information Retrieval," 18(11) *Communications of the ACM* 613.

—— (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, MA: Addison-Wesley.

Schonhardt-Bailey, C. (2005) "Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches," 38(4) *PS: Political Science & Politics* 701.

—— (2006) *From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective*. Cambridge, MA: MIT Press.

Sebastiani, Fabrizio (2002) "Machine Learning in Automated Text Categorization," 34(1) *ACM Computing Surveys* 1.

Shapiro, Martin (1968) *The Supreme Court and Administrative Agencies*. New York: Free Press.

Songer, Donald R. (1989) *U.S. Courts of Appeals Database*. Available at <http://www.as.uky.edu/polisci/ulmerproject/appctdata.htm>.

Songer, Donald R., & Reginald S. Sheehan (1993) "Interest Group Success in the Courts: *Amicus* Participation in the Supreme Court," 46 *Political Research Q.* 339.

Spaeth, Harold J. (2006) *United States Supreme Court Judicial Database*. Available at <http://www.as.uky.edu/polisci/ulmerproject/sctdata.htm>.

Wahlbeck, P. J., J. F. Spriggs, & F. Maltzman (1998) "Marshalling the Court: Bargaining and Accommodation on the United States Supreme Court," 42 *American J. of Political Science* 294.

Yang, Yiming, & Jan O. Pedersen (1997) "A Comparative Study of Feature Selection in Text Categorization," presented at the proceedings of the 14th International Conference on Machine Learning. Nastiville, TN.